

Analysis and Estimation of Customer Survival Time in Subscription-based Businesses

Zakariya Mohammed Salih Mohammed



A thesis submitted in partial fulfilment of the requirement for the degree of Doctor of
Philosophy in the Department of Statistics at the Faculty of Natural Science,
University of the Western Cape

Supervisor: Professor Danelle Kotze

Co-supervisor: Professor Johannes Stefan Maritz

September 2008

KEYWORDS

Customer survival time

Subscription-based businesses

Survival analysis

Churn

Customer mean survival time

Customer-centric approach

Customer equity

Customer lifetime value

Retention rate

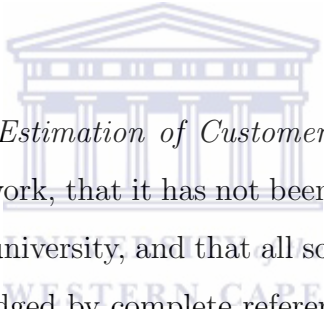
Kaplan-Meier

Extrapolation of the survival curve

Delta method



DECLARATION



I declare that *Analysis and Estimation of Customer Survival Time in Subscription-based Businesses* is my own work, that it has not been submitted before for any degree or examination in any other university, and that all sources I have used or quoted have been indicated and acknowledged by complete references.

Zakariya Mohammed Salih Mohammed

September, 2008

Signed

ABSTRACT

Subscription-based industries have seen a massive expansion in recent decades. In this type of industry the customer has to subscribe to be able to enjoy the service; therefore, well-defined start and end points of the customer relationship with the service provider are known. The length of this relationship, that is the time from subscription to service cancellation, is defined as customer survival time. Unlike transaction-based businesses, where the emphasis is on the quality of a product and customer acquisition, subscription-based businesses focus on the customer and customer retention. A customer focus requires a new approach: managing according to customer equity (the value of a firm's customers) rather than brand equity (the value of a firm's brands). The concept of customer equity is attractive and straightforward, but the implementation and management of the customer equity approach do present some challenges. Amongst these challenges is that customer asset metric - customer lifetime value (the present value of all future profits generated from a customer) - depends upon assumptions about the expected survival time of the customer (Bell *et al.*, 2002; Gupta and Lehmann, 2003). In addition, managing and valuing customers as an asset require extensive data and complex modelling. The aim of this study is to illustrate, adapt and develop methods of survival analysis in analysing and estimating customer survival time in subscription-based businesses. Two particular objectives are studied. The first objective is to redefine the existing survival analysis techniques in business terms and to discuss their uses in order to understand various issues related to the customer-firm relationship. The lesson to be learnt here is the ability of survival analysis techniques to extract important information on customers with regard to their loyalties, risk of

cancellation of the service, and lifetime value. The ultimate outcome of this process of studying customer survival time will be to understand the dynamics and behaviour of customers with respect to their risk of cancellation, survival probability and lifetime value. The results of the estimates of customer mean survival time obtained from different nonparametric and parametric approaches; namely, the Kaplan-Meier method as well as exponential, Weibull and gamma regression models were found to vary greatly showing the importance of the assumption imposed on the distribution of the survival time.

The second objective is to extrapolate the customer survival curve beyond the empirical distribution. The practical motivation for extrapolating the survival curve beyond the empirical distribution originates from two issues; that of calculating survival probabilities (retention rate) beyond the empirical data and of calculating the conditional survival probability and conditional mean survival time at a specific point in time and for a specific time window in the future. The survival probabilities are the main components needed to calculate customer lifetime value and thereafter customer equity. In this regard, we propose a survivor function that can be used to extrapolate the survival probabilities beyond the last observed failure time; the estimation of parameters of the newly proposed extrapolation function is based completely on the Kaplan-Meier estimate of the survival probabilities. The proposed function has shown a good mathematical accuracy. Furthermore, the standard error of the estimate of the extrapolation survival function has been derived. The function is ready to be used by business managers where the objective is to enhance customer retention and to emphasise a customer-centric approach. The extrapolation function can be applied and used beyond the customer survival time data to cover clinical trial applications.

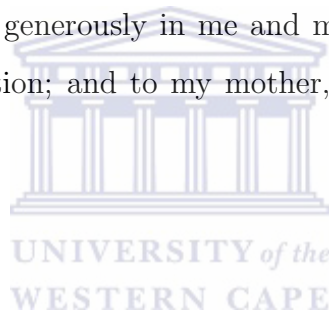
In general the survival analysis techniques were found to be valuable in understanding and managing a customer-firm relationship; yet, much still needs to be done in this area of research to make these techniques that are traditionally used in medical studies more useful and applicable in business settings.

ACKNOWLEDGMENT

I would like to sincerely thank Prof. Danelle Kotze and Prof. Stefan Maritz for their invaluable guidance and support throughout the study. I would also like to thank the Faculty of Natural Science for their financial support. A special word of thanks has to be extended to the academic and administrative staff of the Department of Statistics who made me feel at home and to my family and friends for their support and encouragement. I am thankful for my wife for her love and support. I am sincerely grateful to my primary, intermediate, secondary school teachers as well as university teachers; my current education has only been possible with their efforts. Finally my eternal gratitude to my mother, Arafah Sheikh-Idris; nothing would have been possible without the energy I got from her prayers.

Dedication

This work is dedicated my parents: To the soul of my father, Mohammed Salih, who gave me his blessing and prayers to start my PhD studies just five days before he passed away, and who has invested generously in me and my brothers and sisters, and who has valued and loved education; and to my mother, Arafa Sheikh-Idris, my all-time number one teacher.



Contents

List of Figures	x
List of Tables	xiv
1 Introduction	1
1.1 Background	1
1.2 The purpose of the study	2
1.3 Motivation	4
1.4 Delimitations, limitations and challenges	6
1.5 The structure of the thesis	7
2 Measuring a firm's performance: a customer-centric approach	11
2.1 Customer-centric approach in measuring a firm's performance	12
2.1.1 Emergence of a customer-centric approach in measuring a firm's performance	12
2.1.2 Customer asset metric - customer lifetime value	14
2.2 Survival analysis and customer survival time in measuring a firm's performance	15
2.2.1 The position of customer survival time in the model of measuring a firm's performance	15
2.2.2 Models for customer lifetime value	17
2.2.3 Projection approaches in customer lifetime value models	19
2.3 Summary	20

3 The methods of survival analysis in analysing and estimating customer survival time	22
3.1 Customer survival time: basic approach and formulation	23
3.1.1 Basic approach	23
3.1.2 Basic formulation of customer survival and hazard functions . .	24
3.1.3 Underlying distribution of survival time and customer survival time	25
3.1.4 Survival analysis from medical application to business application	26
3.2 Survival analysis in understanding and maintaining customer-firm relationship	28
3.2.1 Survival analysis and comparison of different marketing campaigns and different customer groups	28
3.2.2 Survival analysis and identification of significant variables that affect customer survival time	29
3.2.3 Survival analysis and estimation of the customer mean survival time	29
3.3 Summary	31
4 Customer survival time data: Application with discussion	33
4.1 The data	34
4.1.1 Data extraction and preparation	34
4.1.2 Data description	35
4.1.3 Approach to data analysis	38
4.2 Identifying the significant variables that affect customer survival time using the Cox model	39
4.3 Analysing the risk of service cancellation using the hazard and the hazard ratio	41
4.3.1 The use of Cox hazard ratios	41
4.3.2 The use of the Nelson-Aalen integrated hazard to understand the risk of cancellation of service over time	44

4.4	Understanding customer survival probabilities using the Kaplan-Meier method	48
4.5	Estimating the customer's mean survival time using both the non-parametric and parametric model	53
4.6	Summary and discussion	57
5	Extrapolation of the survival curve	61
5.1	Motivation	61
5.2	The proposed extrapolation function	62
5.3	The mathematical check of the proposed survivor function	63
5.3.1	Checking accuracy of the extrapolation function	64
5.3.2	Results of the mathematical check	64
5.3.3	Summary of the results of the mathematical check	82
5.4	The derivation of the standard error of the extrapolation function	83
5.5	The conditional survival function	88
5.6	Applications	88
5.6.1	Estimation of future survival probabilities	88
5.6.2	Establishing confidence limits for customer lifetime value	91
5.7	Summary and conclusion	93
6	Conclusion, recommendations and directions for further studies	96
	Bibliography	100
A	Octave code	110
A.1	The main file	110
A.2	Auxiliary functions	114

List of Figures

2.2.1 The position of customer survival time in the process of measuring a firm's performance based on customer-centric approach	16
3.3.1 Survival analysis techniques and the problem of analysing and estimating customer survival time.	32
4.4.1 Customer survival probabilities by gender (95% confidence interval is attached)	49
4.4.2 Customer survival probabilities by age group (95% confidence interval is attached)	49
4.4.3 Customer survival probabilities by language (95% confidence interval is attached)	50
4.4.4 Customer survival probabilities according to usage purposes (95% confidence interval is attached)	51
4.4.5 Customer survival probabilities by gender and in group less than 26 years (95% confidence interval is attached)	52
4.4.6 Customer survival probabilities by gender and in group 26 to 40 years (95% confidence interval is attached)	52
4.4.7 Customer survival probabilities by gender and in age group more than 40 years (95% confidence interval is attached)	53
5.3.1 The theoretical and fitted survival curve for the gamma distribution . .	66

5.3.2 The theoretical and fitted survival curve for the mixture of two exponential distributions (1)	68
5.3.3 The theoretical and fitted survival curve for the mixture of two exponential distributions	70
5.3.4 The theoretical and fitted survival curve for the mixture of two Weibull distributions (1)	72
5.3.5 The theoretical and fitted survival curve for the mixture of two Weibull distributions (2)	74
5.3.6 The theoretical and fitted survival curve for the mixture of two gamma distributions	76
5.3.7 The theoretical and fitted survival curve for the mixture of exponential and Weibull distributions	78
5.3.8 The theoretical and fitted survival curve for the mixture of exponential and gamma distributions	80
5.3.9 The theoretical and fitted survival curve for the mixture of Weibull and gamma distributions	82
5.6.1 Confidence band of the survival probabilities for customers of age less than 26 years	89
5.6.2 Confidence band of the survival probabilities for customers of age between 26 and 40 years	90
5.6.3 Confidence band of the survival probabilities for customers of age more than 40 years	91

List of Tables

4.1.1 The distribution of customers according to their demographic characteristics	37
4.1.2 The distribution of customers according to their usage-related characteristics	38
4.2.1 Identifying the significant variables that affect customer survival using the stratified Cox model	39
4.2.2 The joint p-value calculated for the grouping variables	41
4.3.1 The hazard ratio calculated for the age group in the row compared to the age group in the column (95% confidence interval for the hazard ratio is attached as well).	41
4.3.2 The hazard ratio calculated for language in the row compared to language in the column (95% confidence interval for the hazard ratio is attached as well).	42
4.3.3 The hazard ratio calculated for marketing city in the row compared to marketing city in the column (95% confidence interval for the hazard ratio is attached as well).	44
4.3.4 The cumulative hazard of service cancellation over time by gender (95% confidence interval is attached).	44
4.3.5 The cumulative hazard of service cancellation over time by age group (95% confidence interval is attached).	45

4.3.6 The cumulative hazard of service cancellation over time by language (95% confidence interval is attached).	46
4.3.7 The cumulative hazard of service cancellation over time by WiFi usage (95% confidence interval is attached).	46
4.3.8 The cumulative hazard of service cancellation over time by IT back- ground (95% confidence interval is attached).	47
4.3.9 The cumulative hazard of service cancellation over time by service usage purpose (95% confidence interval is attached).	47
4.5.1 Estimation of customer mean survival time (in months) by gender and type of the model	54
4.5.2 The range of the point estimate of customer mean survival time for different modelling approaches and by gender	54
4.5.3 Estimation of customer mean survival time by age group and type of the model	55
4.5.4 The range of the point estimate of customer mean survival time for different modelling approaches and by age group	55
4.5.5 Estimation of customer mean survival time by usage purpose and type of the model	56
4.5.6 The range of the point estimate of customer mean survival time for different modelling approaches and by usage purpose	56
5.3.1 The theoretical and fitted survival probabilities for the gamma distribution	65
5.3.2 The theoretical and fitted survival probabilities for the mixture of two exponential distributions (1)	67
5.3.3 The theoretical and fitted survival probabilities for the mixture of two exponential distributions (2)	69
5.3.4 The theoretical and fitted survival probabilities for the mixture of two Weibull distributions (1)	71
5.3.5 The theoretical and fitted survival probabilities for the mixture of two Weibull distributions (2)	73


5.3.6 The theoretical and fitted survival probabilities for the mixture of two gamma distributions	75
5.3.7 The theoretical and fitted survival probabilities for the mixture of ex- ponential and Weibull distributions	77
5.3.8 The theoretical and fitted survival probabilities for the mixture of ex- ponential and gamma distributions	79
5.3.9 The theoretical and fitted survival probabilities for the case of the mix- ture of Weibull and gamma distributions	81
5.3.10 The maximum norms for different projection periods and different the- oretical distributions	95



Chapter 1

Introduction

1.1 Background



Subscription-based industries have seen a massive expansion in recent decades. In this type of industry the customer has to subscribe to be able to enjoy the service; therefore well-defined start and end points of the customer relationship with the service provider are known. The length of this relationship is defined to be customer survival time - that is the time from subscription to service cancellation. In business literature, service cancellation is widely known as churn. Unlike in transaction-based businesses, where the emphasis is on the quality of product and customer acquisition, subscription-based businesses focus on the customer and customer retention (Rust *et al.*, 2000; Roberts, 2000). A customer focus requires a new approach: managing according to customer equity (the value of a firm's customers) rather than brand equity (the value of a firm's brands), and focusing on customer profitability rather than product profitability (Rust *et al.*, 2000). Robert Blattberg and John Deighton introduced the term customer equity (Blattberg and Deighton, 1996) which means the total discounted future net revenue that a firm expects from its relationship with its customers today. The metric used to measure customer equity is customer lifetime value, which refers to the total net revenue that a firm expects today from its future relationship with a customer (Gupta and Lehmann, 2003). The concept of customer equity is attractive and straightfor-

ward, but the implementation and management of the customer equity approach do present some barriers (Bell *et al.*, 2002; Gupta and Lehmann, 2003; Shah *et al.*, 2006, Zeithaml *et al.*, 2006). Amongst these barriers is that customer asset metric - customer lifetime value - depends upon assumptions about the expected survival time of the customer (Bell *et al.*, 2002). In addition, managing and valuing the customer as an intangible asset requires extensive data and complex modelling. Moreover a strong link between investment in a customer-centric approach and the firm's financial value needs to be shown to the investors by using a simple numerical measure that is based on the customer-centric approach and is presented in an easy way similar to traditional financial measures such as profit, cash flow and return on investment (Gupta and Lehmann, 2003; Raab, 2007).

1.2 The purpose of the study

The aim of this study is to illustrate, adapt and develop methods of survival analysis in analysing and estimating customer survival time in subscription-based businesses. Two particular objectives are set:

The first objective is to redefine the existing survival analysis techniques in business terms and to discuss their uses in order to understand various issues related to the customer-firm relationship.

In relation to the redefinition of the current survival analysis techniques to meet business needs, the objective is to study:

- The basic formulation of survival and hazard functions in business terms.
- The underlying assumptions about the distribution of customer survival time.
- The differences between survival analysis in the medical field and business field due to the nature of business data and business needs.

- The approach of survival analysis in comparing different marketing campaigns and different customer groups and identification of significant variables that affect customer survival time.

With regard to the use of survival analysis in investigating customer-firm relationship, our objective is to study:

- The use of the Cox model (Cox, 1972) to identify significant variables that affect customer survival time.
- The analysis of the risk of service cancellation using the hazard calculated from the Nelson-Aalen method (Nelson, 1972; Aalen, 1978) and the hazard ratio calculated from the Cox model (Cox, 1972).
- The understanding of customer survival probabilities using the Kaplan-Meier method (Kaplan and Meier, 1958).
- The estimation of the customer mean survival time using both nonparametric and parametric methods. This requires extrapolation of the survival curve.

The second objective is to extrapolate the customer survival curve beyond the empirical distribution. In this regard, the particular aim is to:

- Propose a survival function that can be used to extrapolate the survival probabilities beyond the empirical data for projection purposes based on the Kaplan-Meier estimate of the survival probabilities in the observation period.
- Derive the standard error of the estimate of the proposed extrapolation function.
- Introduce methods to derive the conditional survival probabilities using the extrapolation function.
- Suggest an approach to estimate the standard error of the conditional survival time.

The aim of this study is to make two main contributions.

The first one is to provide a comprehensive understanding of the use of survival analysis in utilising customer-firm relationships. A considerable effort is made to investigate customer survival time in subscription-based businesses. The analysis is related to the evaluation of a firm's performance based on a customer-centric approach. Materials related to this approach of analysis is presented throughout the thesis, but is mainly explored in chapter two, three and four.

The second contribution is the work on the extrapolation of the survival function on the basis of Kaplan-Meier estimates and the estimation of the standard error of the proposed extrapolation survival function. This extrapolation function is readily available for practical use for a reasonable projection period (the latter to be decided by expertise in the field where the model has to be applied) and a given standard error. The codes are written in the open source Octave. The material following this approach of the study is presented in chapter five.

1.3 Motivation

Customer survival time is the main component that has to be estimated in order to calculate customer lifetime value. Customer lifetime value is considered to be the most acceptable and widely used metric to evaluate a firm's performance based on a customer-centric approach (Mani *et al.*, 1999; Lu, 2003; Gupta and Lehmann, 2003). Although some researchers have pointed out the importance of survival analysis in calculating customer lifetime value, little research has been done in this area. Most models for customer lifetime value that have been developed so far are deterministic models; these models have a constant survival probability (retention rate) and an infinite time horizon. It is argued here that a constant customer retention rate (survival probability) and infinite time horizon projections are counter-intuitive. The change in

customer characteristics and behaviour, marketing campaigns, and market dynamics are potential factors that affect retention rate over time. Market dynamics and the change in customer behaviour make it impossible for any model to make good predictions of the distant future based on only current information. Gupta and Lehmann stated that “practically, the retention rate is one of the most difficult metrics to empirically estimate” (Gupta and Lehmann, 2003).

In some subscription-based industries where customers have an almost equal monthly margin (total profit generated over a month), customer survival time becomes an important measure for identifying the most valuable customers; that is the future expected revenue from a customer evaluated at a specific point in time will be proportional to his/her expected future survival time. A typical question here is for how long a customer will be actively engaged in business with the firm. This question could be answered based on individual customer characteristics or based on segment (group of customers) characteristics using survival analysis techniques. It is our objective in this study to analyse, understand and develop models that consider a nonconstant retention rate estimated from the empirical data and make predictions for a reasonable time horizon (the choice of time horizon should be based on the nature of the industry).

The management of the customer-firm relationship, management and implementation of a customer-centric approach and the need to justify investment in customer-centric initiatives have arisen so that the need to investigate the adequacy and applicability of existing models becomes necessary (Zeithaml *et al.*, 2006; Raab, 2007). The usefulness of survival analysis is not limited to the calculation of customer lifetime alone. The survival probabilities can also be used to plan retention and acquisition programs.

Guided by the above-mentioned motivations the aim is to provide a better understanding of the role of survival analysis in emphasising a customer-centric approach and

to suggest a solution to the problem of extrapolation. However, the proposed extrapolation model goes beyond the business setting to clinical trials and cost-effectiveness studies as well. Furthermore, this study is meant to enhance and contribute to the understanding, literature and application of survival analysis techniques.

1.4 Delimitations, limitations and challenges

This study considers customer survival time in subscription-based businesses where the starting point and end point of the customer relationship with the service provider is well-known. The study does not differentiate between a new subscriber or an old customer who has cancelled the service for some time and rejoined the service provider later. In a real business setting the survival time and lifetime value models of customers who activated their service might be different than those of newly joining customers. Customers who rejoin the service provider have a lower cost of retention and acquisition than those in the newly subscribing group. This is due to the difference in their knowledge about and attitude towards the service provider.

It has to be noted that while the mathematical accuracy of the proposed model gave a good indication of the adequacy of the model in general, the application of the model has to be carefully handled (e.g. the number of data points to be taken in order to estimate the parameters of the extrapolation function). This is especially necessary in the presence of extreme right censoring, which is the case in most customer survival time data.

While the importance of customers has been acknowledged for several decades, the use of a customer-centric approach in measuring a firm's performance developed more recently when Robert Blattberg and John Deighton introduced the approach of customer equity (Blattberg and Deighton, 1996). Several attempts since then have been made to model the customer equity metric - customer lifetime value (Berger and Nasr,

1998; Reinartz and Kumar, 2000; Thomas, 2001, Jain and Singh, 2002; Hogan *et al.*, 2002; Berger *et al.*, 2002; Libai *et al.*, 2002; Hogan *et al.*, 2003; Lu, 2003; Gupta and Lehmann, 2003; Gupta *et al.*, 2004; Bolton *et al.*, 2004; Kumar *et al.*, 2004; Campbell and Frei, 2004; Kumar and Petersen, 2005; Gupta *et al.*, 2006, Kumar *et al.*, 2006). However, most of these modelling attempts were based on deterministic approaches in estimating the customer survival time component in the customer lifetime value model and a very limited investigation has been made in the direction of probabilistic approaches. The challenge is, then, one of having limited literature in the use of statistical approaches to deal with customer survival time, particularly in literature using survival analysis techniques to model customer lifetime value and implement a customer-centric approach.

Although the problem of extrapolation is not new for statisticians and mathematicians, it remains a challenging topic and it is very important. In the case of the Kaplan-Meier survival curve extension, not much work has been done, except by Gross and Clark, 1975, Lagakos, 1979, Moeschberger and Klein, 1985 and King *et al.*, 2003.

1.5 The structure of the thesis

The thesis is composed of six chapters. In chapter one we have introduced the problem, stated our objectives and highlighted the motivations behind the choice of this topic. The limitation and the delimitation of the study have also been stated. In chapter two, a review of the literature is presented. The practical value of this study has been discussed here and the business motivations are presented. This is done in three main sections. In the first section it is shown that the customer has emerged as an important intangible asset to be measured in order to evaluate the firm's performance. The customer equity approach is introduced as well as its metric - customer lifetime value. In the second section it is shown that customer survival time is well-positioned in the process of measuring a firm's performance and in implementing customer-centric ap-

proaches. The models that are used to calculate customer lifetime value are reviewed. All the issues related to the objectives of this study that emerged during the course of reviewing the topics of this chapter are highlighted. In the last section the main points of this chapter are summarised.

In chapter three a basic formulation of survival analysis techniques are provided. Unlike the previous chapter where emphasis was on the business of understanding the customer and customer survival, this chapter focuses on explaining the use of survival analysis techniques in tackling various business issues around the customer-firm relationship. It consists of three main sections. In the first section the basic formulation of survival analysis in the customer survival time context is provided and scholars' work with respect to the issue of churn is reviewed. Similarities and differences between application of survival analysis in the medical context and business context are discussed. The second section presents a number of important practical business problems in understanding customer-firm relationships that can be resolved using various survival analysis techniques. The last section aims to summarise this chapter.

Chapter four contains five main sections. In this chapter various survival analysis techniques are used in analysing a particular data set that has been extracted from the database of a company that provides a subscription-based service. The aim is to give a better understanding of customer survival time and to investigate the usefulness of using survival analysis to understand various business questions related to customer survival time in subscription-based businesses. Any methodological implications and challenges facing the current survival techniques emerging in this setting are discussed.

To analyse and understand customer survival time, various approaches to survival time data analysis were applied on a sample of the data set. The analysis was conducted at different levels. On the first level the significant variables that affect customer survival time are identified. For this purpose the results obtained from a stratified Cox

regression are presented. The outcome of this level of analysis is presented in section one. On the second level hazard ratios obtained from the stratified Cox model are used to evaluate the effects of the different levels of each variable. The hazard obtained from Nelson-Aalen is presented to get a better understanding of the change in risk of cancellation over time. The second level of analysis is presented in section two. On the third level of the analysis the change in the survival probability over time using the Kaplan-Meier method is considered. The results of the Kaplan-Meier analysis is presented in section three. On the final level of analysis the estimation of the mean obtained from the Kaplan-Meier method and the parametric method using the exponential distribution, Weibull distribution and gamma distribution are presented and compared, and these results are presented in section four. Methodological and business insights are discussed in each section to get a better understanding of the application and the theory. In the last section of this chapter a brief summary is presented.

Chapter five is considered the most important chapter in this study. The issue of extrapolating the survival curve is investigated beyond the empirical data and is presented taking into account the nature and the context of the business problem and customer survival time. During the research process, this issue has appeared to be a crucial element in building a good model that is based on a customer-centric approach. This chapter has seven main sections. In the first section the motivation behind extrapolating the survival curve are discussed while in the second section a survivor function is proposed to be used beyond the empirical data. The third section investigates the mathematical accuracy of the proposed model and the fourth section is dedicated to the derivation of the standard error of the estimate obtained from the proposed function. In the fifth section the conditional survival time and the conditional mean survival time are presented. In the sixth section an application of the proposed model on a particular data set is presented. In the last section follows a discussion and the chapter is summarised.

The last chapter, chapter six, concludes the study and points out recommendations.

Some materials of this thesis, particularly parts of chapter three and chapter four, were published, as summarised below:


- Mohammed, Z. and Kotze, D. (2005). Survival data mining in the telecommunications industries: usefulness and complications, Data mining, text mining and their business applications, Wessex Institute of Technology Transaction on Information and Communication Technologies, Volume 35: 505-512.
- Mohammed, Z., Maritz, J. S. and Kotze, D. (2007a). Customer survival time in subscription-based businesses (case of Internet service providers). Data mining, text mining and their business applications, Wessex Institute of Technology Transaction on Information and Communication Technologies, Volume 38:303-310.
- Mohammed, Z., Maritz, J. S. and Kotze, D. (2007b). Estimation of the customers' mean survival time in subscription-based businesses. Data mining, text mining and their business applications, Wessex Institute of Technology Transaction on Information and Communication Technologies, Volume 38: 285-292.

The following three articles are in the process of being submitted to scholarly journals:

- On the extrapolation of the Kaplan-Meier survival curve.
- Projecting customer lifetime value and customer equity: A survival analysis approach.
- Notes on the application of survival analysis in the business field.

Chapter 2

Measuring a firm's performance: a customer-centric approach



The objective of this chapter is to review the literature related to the practical value of this study and to present business motivations behind the survival analysis. The chapter has three main sections. In the first section, we show how the customer has emerged as an important intangible asset to be measured in order to evaluate the firm's performance. We introduce the customer equity approach and its metric - customer lifetime value. In the second section, we show how customer survival time is well positioned in the process of measuring the firm's performance and in implementing customer-centric approaches. We also review the models that are used to calculate customer lifetime value. All the issues related to the objective of this study that emerged during the course of reviewing the topics of this chapter are highlighted. In the last section, we summarise the main points of this chapter.

2.1 Customer-centric approach in measuring a firm's performance

2.1.1 Emergence of a customer-centric approach in measuring a firm's performance

In the past three decades the service industry has expanded rapidly creating a new business world economy. The new business world is increasingly focusing on customers rather than products. The old economy was centered on goods, had a transaction-based nature, focused on acquiring customers and product-based thinking while the new economy is service-centered, subscription-based in its nature, focusing on customer retention, and customer-based thinking (Rust *et al.*, 2000; Roberts, 2000). This transformation has made intangible assets - in particular customers - critical to a firm. Managing and measuring performance taking this intangible asset into consideration therefore becomes essential. The usefulness of reported earnings, cash flow and book value in reflecting the value of a firm have been important in recent decades (Lev and Zarowin, 1999). The market value of the top 500 companies in the United States is almost six times the book value (the net value stated on the balance sheet); that is for every six dollars in the market value of a firm only one dollar is represented in the balance sheet (Lev, 2001). In the issue of 22 May 2001 of The New York Times magazine, Floyd Norris wrote: "Intangible assets are, by definition, hard to see and even harder to fix a precise value for. But a widening consensus is growing that the importance of such assets - from brand names and customer lists to trademarks and patents - means that investors need to know more about them" (Norris, 2001).

There is no doubt that customers are not only an essential part of the successful business, but that business can not exist without them (Gouthier and Schmid, 2003). Therefore, decisions made in the business environment have to be centred on customers. A customer focus requires a new approach: managing according to customer equity - the value of a firm's customers - rather than brand equity (the value of a firm's

brands), and focusing on customer profitability rather than product profitability (Rust *et al.*, 2000). Robert Blattberg and John Deighton introduced the term customer equity (Blattberg and Deighton, 1996) and defined it to mean the total discounted future net revenue that a firm expects from its relationship with its customers today.

Although the concept of customer equity is very attractive and straightforward, the implementation and the management of a customer equity approach seems to present some barriers (Bell *et al.*, 2002; Gupta and Lehmann, 2003; Shah *et al.*, 2006). Amongst these barriers is that customer asset metric - customer lifetime value (the total net revenue that a firm expects today from its future relationship with a customer) - depends upon assumptions about the future stream of income from a customer, the appropriate allocation of costs to customers, the discount factor and the expected lifetime of a customer (Bell *et al.*, 2002). Gupta and Lehmann (2003) stated two reasons for the complexity of managing and valuing customers as an intangible asset: firstly, it requires extensive data and complex modelling and secondly a strong link between investment in a customer-centric approach and the firm's financial value needs to be shown to the investors. Marketers have to justify investment for investors and for managers who are outside the marketing department and this should be done in a simple numerical measure that is based on the customer-centric approach and presented in an easy way similar to traditional financial measures such as profit, cash flow and return on investment (Raab, 2007). The new measure should capture the value of customers as an intangible asset. That is why analysts have suggested customer lifetime value to be used as a customer asset metric. In this study we are working on one of the items that have been highlighted by Bell and others (Bell *et al.*, 2002); that is the expected lifetime of a customer. We are also aiming to tackle the issue related to modelling complexity that has been pointed out by Gupta and Lehman (2003) and to face the challenge of making the return on investment in a customer-centric approach accurately measurable. The next section focuses on the customer asset metric, namely, customer life time value.

2.1.2 Customer asset metric - customer lifetime value

Customer lifetime value is defined as the present value of all future profits generated from a customer during his/her relationship with the firm (Gupta and Lehmann, 2003). However, Pfeifer and others have argued that the word value should mean cash flow rather than profit in order to match the word value in finance (Pfeifer *et al.*, 2005). This would make sense if one would like to value the customer as an asset, which is the case in a customer-centric approach. Gupta and others continue with the Pfeifer debate (Gupta *et al.*, 2006). Customer equity is defined in terms of the customer lifetime value as the sum of all customers' lifetime values. The understanding of customer lifetime value has been of great importance to firms in order to go through the process of implementing a customer-centric approach successfully; this has made it a topical area of research (Reinartz and Kumar, 2000; Kumar *et al.*, 2006).

Customer lifetime value has two important uses: firstly, it is used in planning differential marketing initiatives targeting each customer (or segment) and best marketing practices (Kumar *et al.*, 2004); secondly, it is to be used to understand and evaluate the relationship between marketing actions and shareholders' value (Hogan *et al.*, 2002, Rosset *et al.*, 2002).

Several components need to be considered should one want to calculate customer lifetime value; namely, the customer survival time (or customer lifetime), revenue gained from a customer over a unit of time, e.g., per month, discount factor (a discount rate to obtain the present value of cash that will be received in future) and the cost related to maintaining the relationship with the customer. This study focuses on the analysis and the estimation of the first component of the customer lifetime value model i.e. customer survival time.

The next section is concerned with how customer survival is positioned in the process of measuring a firm's performance and calculating the customer lifetime value. Another term that has been used to describe customer survival probability is retention rate.

2.2 Survival analysis and customer survival time in measuring a firm's performance

2.2.1 The position of customer survival time in the model of measuring a firm's performance

Customer survival time is a crucial element in the process of implementing a customer-centric approach, because it is the most important element in calculating customer lifetime value. Lu (2003) emphasised that the customer survival curve and the customer monthly margin are the most important components in modelling customer lifetime value in the telecommunication industries. Estimation of customer survival time can enable managers to test different marketing initiatives and strategies that have been planned to retain customers, and it will allow them to set in place a solid plan for acquiring new customers if it is necessary to do so. Gupta and Lehmann stated that “practically, the retention rate is one of the most difficult metrics to empirically estimate” (Gupta and Lehmann, 2003). It is in this regard that we expect our study to make a contribution. This is achieved by giving a better understanding of the question of customer survival time estimation (and analysis) from the empirical data using existing survival analysis techniques and by proposing a model to extrapolate the survival probabilities beyond the empirical data.

To accommodate the business settings characteristics, survival analysis techniques are expected to accommodate various factors that affect customer survival time. Those factors range from internal factors related to the customer (customer behaviour and characteristics), the internal factors associated with the company (service quality and marketing program) to the external factors (factors associated with competitors' activities and general economic conditions).

In some subscription-based industries, where customers have an almost equal monthly margin, customer survival time becomes an important measure to identify the most valuable customers; that is, the future expected revenue from a customer evaluated at

a specific point in time will be proportional to his/her expected future survival time. A typical question here is for how long a customer will be actively engaged in business with the firm. This is customer survival time. This question could be answered based on individual customer characteristics or based on segment (group of customers) characteristics by calculating the expected survival time.

The position of customer survival time in the model of measuring a firm's performance is summarised in figure 2.2.1 below.

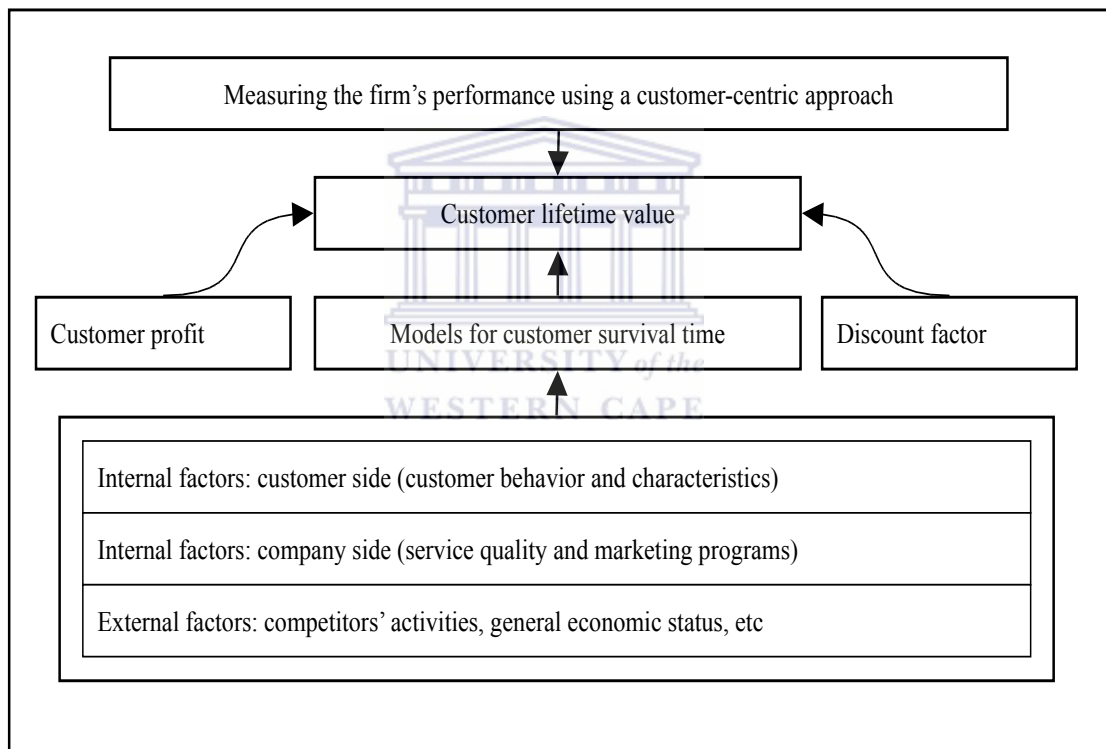


Figure 2.2.1: The position of customer survival time in the process of measuring a firm's performance based on customer-centric approach

In the next section, we review different models for customer lifetime value. Different scenarios on how researchers have dealt with the issue of estimation of customer survival probabilities (or retention rate) over time will be presented.

2.2.2 Models for customer lifetime value

Different models for customer lifetime value have been built over the past 10 years (Berger and Nasr, 1998; Jain and Singh, 2002; Hogan *et al.*, 2002; Berger *et al.*, 2002; Libai *et al.*, 2002; Hogan *et al.*, 2003; Lu, 2003; Gupta and Lehmann, 2003; Gupta *et al.*, 2004; Bolton *et al.*, 2004; Kumar *et al.*, 2004; Campbell and Frei, 2004; Venkatesan and Kumar, 2004; Kumar and Petersen, 2005; Gupta *et al.*, 2006). They share similar objectives as they all try to identify the most valuable customers, obtain the firm's customers' value (customer equity), allocate resources and justify investment in different marketing initiatives. However, to build these models researchers have used different methodologies ranging from fully deterministic mathematical models to statistics and data mining approaches.

Following Gupta and Lehmann (2003), the simplest customer lifetime value (*CLV*) model assuming that we know exactly for how long the customer is going to stay, is

$$CLV = \sum_{t=0}^n \frac{m_t}{(1+d)^t} \quad (2.2.1)$$

where m_t is the margin or a contribution for each customer in a given time t , d is the discount rate and n is the number of time units (time period) over which the customer is assumed to be active. While it is not realistic to assume that we know for certain for how long the customer will remain active, this assumption is relaxed by assuming a particular retention rate (survival probability), that is, a probability of the customer being active in subsequent periods. If we set the probability of a customer being active throughout the period j equal to p_j , then the probability of surviving at the end of period t is

$$r_t = \prod_{j=1}^t p_j.$$

Equation 2.2.1 will become

$$CLV = \sum_{t=0}^n \frac{m_t r_t}{(1+d)^t} \quad (2.2.2)$$

Lu (2003) presented a customer lifetime model for the telecommunication industries

that included the customer margin and customer survival probability (or retention rate). The model is presented as follows:

$$CLV = m \sum_{t=1}^T \frac{S_t}{(1 + d/12)^{t-1}} \quad (2.2.3)$$

where m is the customer monthly margin calculated from the last three months for existing customers and calculated from the last month if the customer is newly acquired, S_t is the series of customer survival probabilities obtained from survival curve, d is the discount rate and T is the number of months for which the customer lifetime value should be calculated.

Mani, Drew, Betz and Datta pointed out that the main challenge in predicting customer lifetime value is the estimation of the customer survival time component (Mani *et al.*, 1999). They argued that the classical survival analysis techniques (such as proportional hazard models) may not work well because of the assumption they make about the linear effect of the covariates, the proportionality assumption and their inability to detect segments of customers whose survival time covariates vary a lot. They proposed a new model that incorporates both the proportional hazard approach and neural networks techniques. They argue that data mining techniques may complement the classical survival techniques.

Figini, Giudici and Brooks used a Bayesian approach to estimate customer lifetime value; in particular, they were interested in customer features selection using a Bayesian approach to model customer lifetime value (Figini *et al.*, 2007). Figini and Giudici made comparisons between classical models such as logistic regression and survival analysis techniques to model customer churn (Figini and Giudici, 2007a). They advocated the use of the Bayesian survival approach rather than the classical survival approach in modelling customer lifetime value, especially estimation that involves a large number of variables to be used (Figini and Giudici, 2007b). Figini investigated survival analysis estimation that is based on the classical partial likelihood. She proposed a Bayesian extension for survival models based on penalised likelihood estimation

(Figini, 2007).

2.2.3 Projection approaches in customer lifetime value models

In relation to the context of our study, two major differences in the approach of tackling customer lifetime value projection models exist in the literature. The distinction between the two approaches is based on the time horizon that they consider when the customer survival time is estimated and on how to estimate customer survival probabilities. The first difference is whether to use an infinite time horizon (Gupta and Lehmann, 2003; Gupta *et al.*, 2004; Gupta *et al.*, 2006) or a finite time horizon (Kumar and Reinartz, 2000, Thomas, 2001) and the second one is whether to take a constant retention rate or a retention rate that varies over time (Gupta *et al.*, 2006);

A number of researchers have assumed a constant customer survival probability over time (Gupta and Lehmann, 2003; Gupta *et al.*, 2004; Gupta *et al.*, 2006). To illustrate this we present an example where we have a constant retention rate of 80%. Take a customer who subscribed to a subscription-based firm such as a telecommunication company, the banking sector or an internet service provider. This constant retention rate of 80% means that the probability of having a customer still active at the end of 5th period (e.g. 5th year) will be equal to 0.33. It is easy to see that having a constant retention rate is neither intuitive nor built on solid theoretical or practical ground. In many industries if the customer could be retained in the initial period of his subscription to the service, the likelihood of developing loyalty becomes high. It is advisable that survival probability be estimated from the empirical data.

Researchers have not agreed on the length of a time horizon over which the projection of customer lifetime value should be made. Many researchers have used an infinite time horizon (Gupta and Lehmann, 2003; Gupta *et al.*, 2004; Fader *et al.*, 2005; Gupta *et al.*, 2006). The infinite time horizon eases the calculation; this could be considered as one of the advantages of this scenario. However, we are not in favour of an infinite time horizon scenario. From a methodological point of view, an infinite time horizon

means more approximations that may produce more errors and predictions that are unreliable and unrealistic. From a business perspective, the fast change in customer behaviour and the market dynamic could make it impossible for statistical models to predict the distant future satisfactorily. It is for these reasons that we favour working with a finite time horizon and a conditional survival time, where the prediction is made on a specific number of time units (e.g., months or years depending on the nature of the industry) and conditioned on information at the time when the prediction has to be made.

2.3 Summary

This chapter has focused on business literature and research on the customer as an intangible asset to be measured and valued. The importance of the customer arises from the domination of the service industry, and the increase in types of businesses that are subscription-based in nature where customer relations have to be managed and maintained. Customer survival time appears to be one of most important components that has to be obtained in order to calculate customer asset metrics - customer lifetime value - and to implement a customer-centric approach in measuring a firm's performance. Researchers acknowledge the difficulty of calculating customer survival probability empirically. Although a few researchers have pointed out the usefulness of survival analysis in calculating customer lifetime value, little research has been done in this area; most models of customer lifetime value that have been developed so far are deterministic models. These models have been dominant with a constant survival probability (retention rate) and an infinite time horizon. We argue that a constant customer retention rate (survival probability) and infinite time horizon projections are counter intuitive. The change in customer characteristics and behaviour, marketing campaigns, and market dynamics are potential factors that affect retention rate over time. Market dynamics and the change in customer behaviour make it impossible for any model to make good predictions of the distant future based on the current infor-

mation. It is our objective in this study to analyse, understand and develop models that consider a nonconstant retention rate estimated from the empirical data and make predictions for a reasonable time horizon (the choice of time horizon should be based on the nature of the industry). Therefore, we motivate further investigation of survival analysis and its use in understanding and modelling customer survival time; hence, customer lifetime value. In particular, we are in favour of estimating the conditional survival probabilities that consider the past customer survival time and the present information on the customer for a reasonably useful projection period.



Chapter 3

The methods of survival analysis in analysing and estimating customer survival time



The objectives of this chapter are to provide a basic formulation of survival analysis techniques when the analysis of customer survival time is considered and to present different problems that survival analysis can solve. Unlike the previous chapter where emphasis was on the business of understanding the customer and customer survival, this chapter focuses on explaining the use of survival analysis techniques in tackling various business issues around the customer-firm relationship. The chapter consists of three main sections. In the first section, we provide the basic formulation of survival analysis in the customer survival time context. We also review scholars' work with respect to the issue of churn. We discuss the similarities and differences between application of survival analysis in the medical context and the business context. The second section presents a number of important practical business problems in understanding customer-firm relationships that can be resolved using various survival analysis techniques. The last section aims to summarise this chapter.

We have published part of the work presented in this chapter in Mohammed and Kotze (2005).

3.1 Customer survival time: basic approach and formulation

3.1.1 Basic approach

Survival analysis concerns time-to-event data analysis (Cox and Oakes, 1984; Oakes, 2001; Klein and Moeschberger, 1997; Hosmer and Lemeshow, 1999; Hougaard, 2001; Lee and Wang, 2003; Lawless, 2003; Jenkins, 2005). In medicine time-to-event can be the time to a certain symptom.

Typically time to event data are censored. Suppose that we followed a group of individuals over a certain period to record the occurrence of the event under study, for instance a heart attack of a patient. By the end of the follow-up period not all individuals will have experienced the event. The time recorded for an individual who has not experienced the event is called a censored time. The specific feature of survival techniques is that they use the data of all available information from both uncensored and censored cases. Another concept is truncation (Cox and Oakes, 1984; Klein and Moeschberger, 1997; Hosmer and Lemeshow, 1999). In addition to censoring, time-to-event data can be truncated. Truncation happens when the investigator makes the starting point of an observation only when it experiences a certain event or satisfies a specific condition. Survival analysis techniques deal with the analysis and modelling of time-to-event data with censoring and truncation.

Subscription-based businesses such as cellular networks, internet service providers, banking services and insurance firms are examples of cases where a well-defined start and end point of the customer's relation with the firm can be found. The time from subscription to the cancellation of a service can be modelled using survival analysis techniques. To study the survival time of a group of customers by recording the date of subscription to the service and date of cancellation of the service, we are likely to have censoring (and truncation) as not all customers will have cancelled the service by the end of the follow-up period.

3.1.2 Basic formulation of customer survival and hazard functions

Let us denote customer survival time by the random variable T . Assuming that T is a continuous random variable, let $F(t) = \text{Prob}(T \leq t)$ be the distribution function of T , and $f(t) = \frac{\partial F(t)}{\partial t}$ be the probability density function of T ; t is a specific value of T . The survival function, $S(t)$, indicates the probability that T exceeds t ; that is:

$$S(t) = \text{Prob}(T > t) = 1 - F(t) \quad (3.1.1)$$

An important concept is the hazard function (or the hazard rate) $h(t)$. It expresses the instantaneous risk of experiencing the event of cancelling the service at time t given that the customer survived until t . It is defined as follows:

$$h(t) = \lim_{\Delta t \rightarrow 0} \left(\frac{1}{\Delta} \text{Prob}(t \leq T < t + \Delta \mid t \geq T) \right) \quad (3.1.2)$$

From equation 3.1.2, it follows that

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\partial S(t)}{\partial t} / S(t) = -\frac{\partial \ln S(t)}{\partial t} \quad (3.1.3)$$

The survival function $S(t)$ can be expressed in terms of $h(t)$ or the cumulative hazard $H(t)$ as follows:

$$S(t) = \exp \left(- \int_0^t h(u) du \right) = \exp(-H(t)) \quad (3.1.4)$$

where $H(t) = \int_0^t h(u) du$ is the accumulation of the hazard over time. The hazard function can be decreasing (customers become more loyal over time), increasing (customers tend to cancel the service over time) or constant (the risk of a customer leaving the service provider does not change over time).

3.1.3 Underlying distribution of survival time and customer survival time

Survival analysis techniques comprise different statistical techniques ranging from non-parametric and semi-parametric to fully parametric methods of analysis and modelling. The choice of the method is largely dependent on the underlying assumption about the distribution, the extent to which we know about the behaviour of the system under study and the type of questions to be investigated. The Kaplan-Meier method (Kaplan and Meier, 1958) gives a nonparametric estimate of the survival function. It applies to homogeneous groups and we can use it to compare survival probabilities across different groups. The Nelson-Aalen method (Nelson, 1972; Aalen, 1978) is a nonparametric technique used to estimate the hazard rate. When parametric analysis and modelling are considered several distributions are available; amongst them are the exponential, Weibull and gamma (Gross and Clark, 1975; Klein and Moeschberger, 1997; Hosmer and Lemeshow, 1999; Lee and Wang, 2003; Lawless, 2003). Each distribution assumes a specific behaviour and shape for the risk of cancelling the service, e.g., constant hazard, monotonically increasing hazard and monotonically decreasing hazard. The advantage of the nonparametric models is that they avoid error arising from misspecifying the underlying distribution. The drawback is that it is much more difficult to report on nonparametric estimates whereas it is easy to do so in parametric models by describing a few parameters. A very popular semi-parametric model in survival analysis is the Cox proportional hazard model or simply Cox model (Cox, 1972; Cox, 1975). The Cox model is defined as follows:

$$h(t) = h_0(t) \exp(\underline{\beta}^T \underline{X}) \quad (3.1.5)$$

where $h_0(t)$ is a baseline hazard function common to all individuals and $\underline{\beta}$ is a vector of regression parameters of dependence of the survival time distribution on the covariate vector \underline{X} . The most interesting feature of the model is its ability to test and estimate the effect of a set of covariates on the hazard rate without making any assumption

about the distribution of the survival data. However, this model is based on the assumption that the hazards are proportional (Anderson, 1982). That is, for a given two observations with different values for the independent variables, the ratio of the hazard functions for those two observations is independent of time. Several methods, such as Schoenfeld and scaled Schoenfeld residuals (Schoenfeld, 1982; Grambsch and Therneau, 1994), have been developed to test this assumption. There are also several tests that can be applied to check the overall specification correctness of the Cox model. Among them, is the link test. The basis of this test is to verify if the coefficient of the squared linear predictors is insignificant; this guarantees that the model is correctly specified and only (and all) the relevant variables are included in the model (Stata, 2003). Several advances were added to the original formulation of the Cox proportional hazard models to accommodate different baseline hazard functions (stratification) and time varying covariates (Therneau and Grambsch, 2001, Lee and Wang, 2003).

The hazard function can be used to study churn. Most marketing initiatives are planned to minimise churn (or maximise retention rate). Losing a customer is a complicated issue that companies would like to avoid. This is not because of the high cost of acquiring a new customer only, but due to the cost of negative word of mouth as well (Hogan *et al.*, 2003; Wangenheim, 2005). However, the customer-centric approach does not only consider minimising churn, but it incorporates the idea of maximising the profit that could be gained from the customer over his survival time period.

3.1.4 Survival analysis from medical application to business application

While the main role of survival analysis in medical applications is to identify influential factors that affect the life of patients, it has two important functions in the business field. The first is to study factors that can prolong the customer's relation with the firm; this is a similar application to the main object of survival analysis in the medical field. In business terminology these techniques can effectively be used to test the impact of marketing campaigns (e.g. assessing the effectiveness of different retention

programs, different levels of a campaign, different incentives being used to upgrade old customers to be more profitable and to add on new products or services). Secondly, is to estimate the expected survival time of a customer. This is the most important difference in emphasis between the two fields of application, because the expected survival time is intimately connected with the expected revenue. This is helpful to identify and target the most profitable customers and to evaluate the firm's value by taking into consideration a customer approach. The nature of customer survival time and objectives (roles) of survival analysis in business require a careful handling of the problem. The nature of the problems that come from the business environment involves several serious issues that can affect the applicability and the implementation of the survival analysis techniques. Firstly, both the customer and company can initiate the termination of a business relationship. The company can initiate the termination of the relationship, for example, if the customer is not able to pay. A customer can also leave the company due to unsatisfactory service received. Secondly, multi-cancellation and multi-reactivation have a great impact both on the methodologies used to estimate the survival probability and the conceptual business bases in understanding the cost and profitability of the old customers who reactivated their service. Thirdly, the nature of covariates that lead to a termination of business relations is important, e.g. type of payment (credit card, cash, bank account), contract (existing or not, whether there is any penalty on the departing customers), promotion, and emergence of a new service provider. It is easy to see that each of these covariates can lead to a sudden loss of a considerable number of customers on a specific date, which will result in a survival curve that is non-smooth with a sharp drop at specific dates representing events like the end of a contract, end of promotion and emergence of a new service provider (Linoff, 2004). Finally, because one of our business aims is to identify profitable customers, segment-based survival curves have to be studied.

3.2 Survival analysis in understanding and maintaining customer-firm relationship

3.2.1 Survival analysis and comparison of different marketing campaigns and different customer groups

The study of customers based on their segments has been highly motivated in business literature, especially in industries where service providers have multimillion subscribers (Alfansi and Sargeant, 2000; Driver and Johnston, 2001; Libai *et al.*, 2002; Weinstein, 2002; Andronikidis and Dimitriadis, 2003; Badgett and Stone, 2005). In such firms it will not be viable for the marketing department to follow the customer on an individual basis. The marketers would rather look at categories of similar behaviours or similar reactions. Motivated with these types of business views, the Kaplan-Meier (Kaplan and Meier, 1958) and Nelson-Aalen (Nelson, 1972 and Aalen, 1978) methods are of great benefit to understand customer survival time without making any assumption about the distribution of customer survival time.

One of the important uses of the Kaplan-Meier non-parametric method is in assessing the differences between survival chances of different groups. Two typical examples are assessing the differences between different levels of marketing campaigns on prolonging customer survival time (enhancing retention rate) and studying the behaviour of different customer groups based on variables such as demographic variables or service usage related variables.

Several tests such as the log-rank test (Mantel, 1966; Peto and Peto, 1972), are applied to test the equality of the survival curves of different groups. If the differences between the survival curves appeared to be statistically significant, then the survival curves (together with a confidence band) produced by Kaplan-Meier method are plotted to see when the differences occur and how big the differences are.

Should a firm want to investigate which of the groups of customers are likely to cancel the service earlier, then the hazard of cancelling the service can be calculated

from Nelson-Aalen. While the Kaplan-Meier method can be used to see which group is more loyal, the Nelson-Aalen method can be used to check which group is more likely to leave the service provider.

3.2.2 Survival analysis and identification of significant variables that affect customer survival time

The Cox proportional hazard model (Cox, 1972; Cox, 1975) is frequently used in medical studies to investigate the relationship between the survival of a patient and a set of explanatory variables. The model provides an estimate of the effect of the intervention (or a characteristic) on patient survival after adjusting for other explanatory variables, and estimates the hazard (or risk) of having the occurrence of the event of interest for individuals given the values of their explanatory variables (Walters, 2001).

This model can be applied to business problems. In many situations the marketing department is interested in investigating the factors that affect customer survival and cancellation probabilities. These factors can be customer characteristics such as the demographic profile or marketing initiatives that have been made to retain customers. The Cox model will help to identify factors that are significant in predicting customer survival (or cancellation of service) and then to plan accordingly. It helps also in studying the significance of any marketing campaign that has been focused on customers so as to justify any investment towards establishing a customer-centric culture in the firm.

3.2.3 Survival analysis and estimation of the customer mean survival time

A typical question of great importance is for how long a specific customer is expected to stay active with a company before switching to another service provider. The mean survival time of a customer that belongs to a segment j , with a segment survival

function $S_j(t)$, is μ_j , where

$$\mu_j = \int_0^{\infty} S_j(t) dt \quad (3.2.1)$$

Although it is not certain that the customer who stays longer is the more profitable customer, the marketing department - after knowing that this customer is likely to stay longer - can design plans that focus on maximising the profit gained from this loyal customer (for example by upgrading his/her level of usage of the service, up-selling and cross-selling). In some subscription-based industries where customers pay equal contributions to enjoy the service, the customer mean survival time is the only measure of profit.

The calculation of customer mean survival time using an infinite time horizon has both methodological and practical limitations; methodologically it means more approximations may result in unreliable predictions and, from a business perspective, market dynamics and fast changes of customer behaviour make the prediction of the distant future hard to be reasonably accurate. It can be argued that a model that captures the available information and makes a projection to a reasonable time horizon is more desirable than infinite time horizon predictions. The most reasonable time horizon to use should take into account the industry environment where the model is built. It is in this regard that one can use survival analysis to build a conditional mean survival time for a customer or for a group of customers (customers segment). This conditional model should use the past customer survival time and the current characteristics of the customer to make a prediction of the future survival probability for a reasonable projection period. Suppose that we want to calculate the conditional density function and survival functions for a customer who survived up to time t_0 . The conditional probability density function of T given $T > t_0$ is

$$\frac{f(t)}{S(t_0)}$$

for $t \geq t_0$. The conditional survival function is

$$\frac{S(t)}{S(t_0)}$$

for $t \geq t_0$. This means that the conditional expectation of T given $T > t_0$ is

$$\mu_c = t_0 + \frac{1}{S(t_0)} \int_{t_0}^{\infty} S(t) dt \quad (3.2.2)$$

Instead of calculating the conditional expectation from t_0 to infinity, one can choose a reasonable forecasting period, say, from t_0 to $t_0 + \Delta$. Then 3.2.2 will be

$$\mu_c = t_0 + \frac{1}{S(t_0)} \int_{t_0}^{t_0+\Delta} S(t) dt \quad (3.2.3)$$

This can be illustrated using an exponential survival time function, $S(t) = \exp(\frac{-t}{\mu})$, with $E(t) = \mu$. In this case the conditional mean survival time over the period t_0 to $t_0 + \Delta$ will be

$$\mu_c = t_0 + \int_{t_0}^{t_0+\Delta} \frac{S(t)}{S(t_0)} dt = t_0 + \int_{t_0}^{t_0+\Delta} e^{\left(\frac{-t+t_0}{\mu}\right)} dt = t_0 + \mu$$

The choice of Δ is dependent on the industry type and the problem at hand; it can be months (6 months, 12 months ...) or years (1 year, 2 years, 5 years, ...).

3.3 Summary

This chapter has presented the basic formulation of survival analysis in the customer context. We have shown the applicability of survival analysis techniques, including non-parametric, semi-parametric and parametric models, in understanding and analysing various issues related to the customer-firm relationship. These issues are mainly modelling and understanding customer loyalty and churn, comparing different marketing campaigns and different customer groups, identifying the significant variables that affect customer relationship with the firm, and the individual's and segment-based estimation of customer survival time as well as customer conditional survival time.

These issues are summarised in figure 3.3.1. Both the previous chapter and this chapter have considered the projection issue as one of the most important issues in business in general and, especially, in predicting customer survival time. This necessitates that we dedicate considerable attention to the problem of extrapolating the survival curve beyond the empirical distribution using the available data and this will be the business of chapter five.

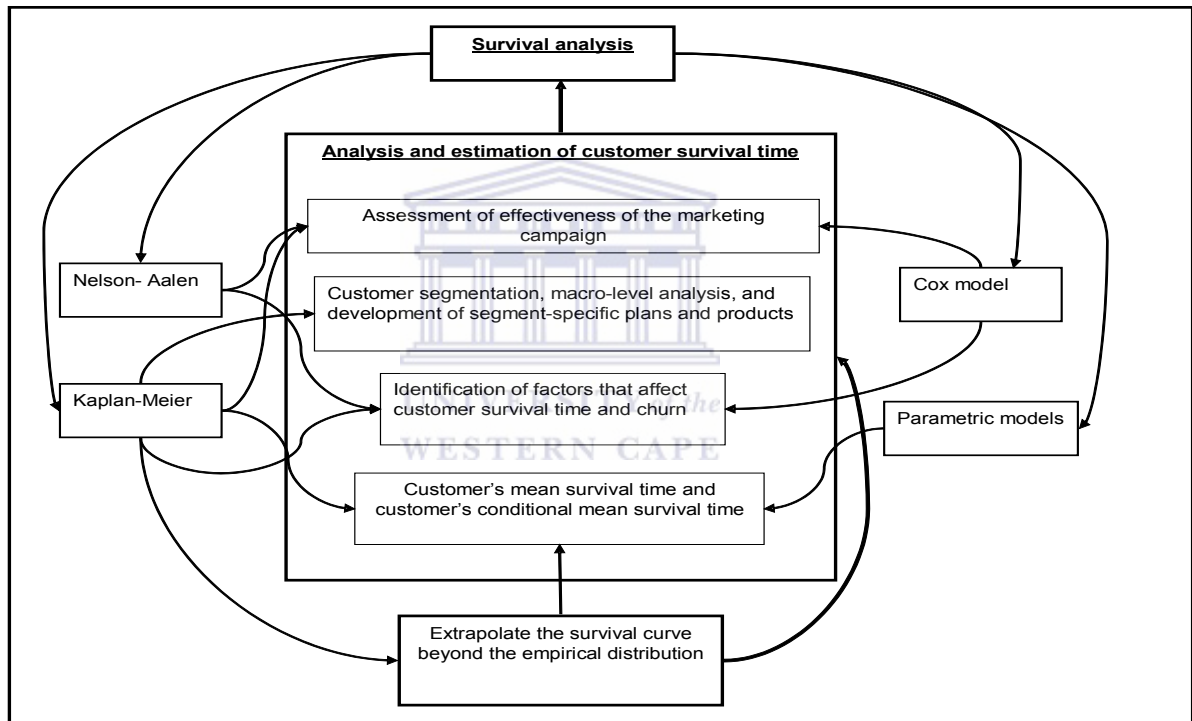



Figure 3.3.1: Survival analysis techniques and the problem of analysing and estimating customer survival time.

In the next chapter, we implement various survival analysis techniques using a particular data set that has been extracted from the database of a company that provides a subscription-based service. We aim to gain a better understanding of the application of these techniques on business data, investigate the usefulness of using survival analysis to understand various business questions related to customer survival time and highlight possible methodological challenges facing the current survival techniques.

Chapter 4

Customer survival time data: Application with discussion



In this chapter, we employ various survival analysis techniques on a particular data set that has been extracted from the database of a company that provides a subscription-based service. The aim of this chapter is to give a better understanding of customer survival time and investigate the usefulness of using survival analysis to understand various business questions related to customer survival time in subscription-based businesses; therefore, any methodological implications and challenges facing the current survival techniques in this setting will be discussed.

To analyse and understand customer survival time, different approaches to survival time data analysis were applied using a sample of customers that had been selected from a database of a subscription-based company. The analysis was conducted on different levels. On the first level, the significant variables that affect customer survival time are identified. For this purpose the results obtained from a stratified Cox regression are presented. The outcome of this level of analysis is presented in section one. On the second level hazard ratios obtained from the stratified Cox model are used to evaluate the effects of the different levels of each variable such as age group, language and marketing city. We also present the hazard obtained from the Nelson-Aalen method to get a better understanding of change in risk of cancellation over time. The second level

of analysis is presented in section two. On the third level of the analysis the change in the survival probability over time using the non-parametric method of Kaplan-Meier is considered; the results are presented in section three. On the final level of analysis the estimation of the mean obtained from the nonparametric method of Kaplan-Meier and the parametric method using the exponential distribution, Weibull distribution and gamma distribution is presented and compared. These results are presented in section four. Methodological and business insights are discussed in each section to get a better understanding of the application and the theory. In the last section of this chapter, a brief summary is presented.

4.1 The data

4.1.1 Data extraction and preparation

The data have been taken from a database of a well-established subscription-based service provider. In this company, a customer can be subscribed at any time point and should notify the company to cancel the service. Hence, the date of subscription to the service and cancellation of the service (if they did cancel) will be exactly known and recorded in the company database.

The data were extracted from the database of the company at the end of the day of 31st July 2005 and before the working hours of the day of 1st August 2005. Three main files of customer information were extracted; a file containing customer survival time data, a file containing demographic data and another file for usage-related data. The three files were extracted in plain text format (.txt) and then imported in SPSS 13.0.

The variable “customer ID” has been used as the primary key (which is unique and exists in all files) to merge all data files (customer survival time data, demographic data and usage-related data) into a single data file with a single record for each customer. The total number of records was 243333. A simple random sample of 30000 customers was selected. Out of this sample 4850 (16.2%) cancelled the service and 25150 (83.8%)

were still active by time we extracted the data.

4.1.2 Data description

The data set contains the following information:

Customer's survival time data

Customer survival time data contains four variables. They are date of subscription, date of cancellation, customer survival time and status. The variable “date of subscription” records the date when a customer subscribed to the service; the earliest possible date of subscription is 1 January 1997 and the last possible date of subscription is 31 July 2005. The variable “status” is recorded as zero, if the customer is still active (censored cases), and one, if the customer has cancelled the service (event). The date when the customer cancelled the service is recorded in the variable “date of cancellation”. Customer survival time is then calculated in the variable “customer survival time”, which is equal to the date of cancellation minus the date of subscription for customers who experienced the event of service cancellation or date of data extraction minus the date of subscription for censored cases. The customer survival time is recorded in months.

Customer's demographic data

The demographic data included gender, date of birth, age, language and marketing city. The variable “gender” is coded to be zero for male and one for female. The variable “age” stores customer age on the day of subscription in years and is equal to date of subscription minus date of birth. The variable “age group” is formed by categorising the customer according to their age into three age groups; the young customers (age less than 26 years: coded as one), the middle age customers (26-40 years: coded as two), and the older customers (more than 40 years: coded as three). Two indicator variables were created from the variable “age group”. The first one is “young customer” (equal to one if the age of the customer is less than 26 and equal to zero if

not) and the second indicator variable is “middle age customer” (equal to one if the age of the customer lies in the closed interval 26 to 40 and equal to zero if not). Customers choose their preferred service language. We have three categories for language: English (coded as one), Afrikaans (coded two), and others (coded three). Two indicator variables were created from the variable “language”. The first one is “English” (equal to one if customer’s preferred service language is English and zero if not) and the second indicator variable is “Afrikaans” (equal to one if customer’s preferred service language is Afrikaans and zero if not). The variable “direct marketing city” consists of the number of cities where the customer has subscribed. This variable has six categories: Cape Town (coded one), Durban (coded two), Johannesburg (coded three), Pretoria (coded four), Witwatersrand (coded five) and other (other small cities: coded six). Five indicator variables were created from the variable “direct marketing city”: marketing city Cape Town (equal to one if the customer’s marketing city is Cape Town and zero if not), marketing city Durban (equal to one if the customer’s marketing city is Durban and zero if not), marketing city Johannesburg (equal to one if the customer’s marketing city is Johannesburg and zero if not), marketing city Pretoria (equal to one if the customer’s marketing city is Pretoria and zero if not), and marketing city Witwatersrand (equal to one if the customer’s marketing city is Witwatersrand and zero if not). The distribution of customers according to their demographic characteristics is provided in table 4.1.1. The missing data are indicated where applicable.

Table 4.1.1: The distribution of customers according to their demographic characteristics

Variable		Frequency (%)
Gender	Male	17669 (58.9%)
	Female	8262 (27.5%)
	missing	4069 (13.6%)
Age group	Less than 26 years	1965 (6.5%)
	26-40 years	8249 (27.5%)
	More than 40 years	8159 (27.2%)
	missing	11627(38.8%)
Language	English	28244 (94.1%)
	Afrikaans	1351 (4.5%)
	other	25 (0.1%)
	missing	380 (1.3%)
Direct marketing city	Cape Town	5986 (19.9%)
	Durban	2271 (7.6%)
	Johannesburg	5883 (19.6%)
	Pretoria	3080 (10.3%)
	Witwatersrand	4249 (14.2%)
	other cities	7083 (23.6%)
	missing	1448 (4.8%)

Customer's usage-related data

This includes three variables, namely, IT background, WiFi usage, and segment. The variable "IT background" records whether the customer has an IT background or not; this variable is coded one if the customer has an IT background and zero if not. The variable "WiFi usage" shows whether the customer uses WiFi (coded one) or does not use WiFi (coded zero). The variable "segment" shows whether the customer uses the service for private purposes only (coded one) or private and business purposes (coded zero). The distribution of customers according to their usage-related characteristics is provided in table 4.1.2. The missing data are indicated where applicable.

Table 4.1.2: The distribution of customers according to their usage-related characteristics

Variable		Frequency (%)
IT background	Yes	28900 (96.3%)
	No	1048 (3.5%)
	missing	52 (0.2%)
WiFi usage	Yes	29803 (99.3%)
	No	197 (0.7%)
Usage purpose segment	Private purposes	11304 (37.7%)
	Private and business puposes	16685 (55.6%)
	missing	2011 (6.7%)

4.1.3 Approach to data analysis

Various survival analysis techniques are applied to the data; they range from non-parametric to semi-parametric and parametric. Two motivations were behind our use of a wider range of survival analysis techniques. Our first motivation is to reach a better understanding of customer survival time. We believe that different survival techniques are suitable for different business questions; therefore, we explore the contribution of these techniques in enhancing the understanding of customer survival time. The second motivation is to find out if the nature of the customer survival data can challenge the existing survival techniques in some aspects. To answer this question we will discuss the extent to which the current models can explain and express the customer survival time. In the rest of this chapter we present the results with discussion obtained from applying various techniques.

Microsoft Excel, SPSS 13.0 (SPSS, 2003) and SPSS 14.0 (SPSS, 2005) were used for data preparation. Data analysis was carried out using Stata 8.0; some graphs were produced in Microsoft Excel from data generated in Stata 8.0 (Stata, 2003).

4.2 Identifying the significant variables that affect customer survival time using the Cox model

Stratified Cox regression was used to identify the significant variables. The stratification is based on the variable "usage purpose". This variable does not satisfy the proportionality assumption; therefore, the variable was used as stratification variable. All the variables included in the model satisfy the proportionality assumption.

Table 4.2.1: Identifying the significant variables that affect customer survival using the stratified Cox model

The variable	Hazard ratio (95% CI)	P-value
Gender (Female)	1.13 (1.05 , 1.23)	0.002
Age group: Less than 26 years	3.81 (3.43 , 4.23)	0.0001
Age group: 26 to 40 years	1.65 (1.52 , 1.80)	0.0001
Language: English	0.69 (0.17 , 2.77)	0.603
Language: Afrikaans	0.98 (0.24 , 3.95)	0.975
Marketing city: Cape Town	1.01 (0.90 , 1.13)	0.860
Marketing city: Durban	1.11 (0.96 , 1.29)	0.157
Marketing city: Johannesburg	0.89 (0.79 , 1.00)	0.050
Marketing city: Pretoria	1.07 (0.94 , 1.21)	0.316
Marketing city: Witwatersrand	1.14 (1.02 , 1.28)	0.023
IT background	0.17 (0.02 , 1.21)	0.076
WiFi usage	0.42 (0.23 , 0.76)	0.004

In table 4.2.1, the hazard ratio (with a 95% confidence interval) and the corresponding p-value for each variable in the stratified Cox regression model are presented. In this table, for each variable with more than two categories (such as age group, language and direct marketing city) the hazard ratio is calculated for each indicator variable that has been created from this variable. Then, the joint p-value for each categorical variable (with more than two categories) is presented in table ???. The results showed that the hazard of a female cancelling the service is 1.13 times the hazard of a male with 95% confidence limits of (1.05, 1.23) and the corresponding p-value is 0.002.

With regard to age group, we compared young customers (age less than 26 years) and middle age customers (age from 26 to 40 years) with the old age group customers

(that is the age group of more than 40 years is used as reference). The hazard of a young age group customer cancelling the service is 3.81 times the hazard of an old age group customer with 95% confidence limits of (3.43, 4.23) and a p-value equal to 0.0001. The hazard of a middle age group customer cancelling the service is 1.65 times the hazard of the old age group customer with 95% confidence limits of (1.52, 1.80) and a p-value equal to 0.0001. In table 4.2.2, the joint p-value for age group is presented and is equal to 0.0001.

The hazard ratios were calculated for English and Afrikaans and compared to the category of “other” (other languages). Customers who chose English as their service language when they subscribed have 0.69 times the hazard of cancelling the service compared to those who chose other languages (with 95% confidence interval (0.17, 2.77) and p-value equal to 0.975). Customers who chose Afrikaans as their service language when they subscribed have 0.98 times the hazard of cancelling the service compared to those who chose other languages (with 95% confidence interval (0.24, 3.95) and p-value equal to 0.603). In table 4.2.2, the joint p-value for the language factor is presented and is equal to 0.0001.

The factor of marketing city was classified into six categories. These categories were Cape Town, Durban, Johannesburg, Pretoria, Wits and others (other cities in South Africa). The hazard ratio for each marketing city compared to the category of “other” (other cities in South Africa) was calculated. Customers from the marketing city of Witwatersrand have the highest risk of cancelling the service when compared to the category of other cities (1.14 hazard ratio). This is followed by customers from the city of Durban (1.11 hazard ratio). In table 4.2.2, the joint p-value for marketing city factor is presented and is equal to 0.003.

A customer with an IT background has 0.17 times the risk of cancelling the service compared to a customer with no IT background; with a 95% confidence interval (0.02, 1.21) and a p-value equal to 0.076. Customers using WiFi have 0.42 times the risk of cancelling the service compared to customers who do not use WiFi (with a 95% confidence interval (0.23, 0.76) and a p-value equal to 0.004).

An in-depth analysis of customer risk of service cancellation using the Cox regression model and the Nelson-Aalen method is presented in the following section.

4.3 Analysing the risk of service cancellation using the hazard and the hazard ratio

4.3.1 The use of Cox hazard ratios

Further analysis of the effect of age group, language and marketing city using the hazard ratios calculated from Cox regression is presented in this subsection. In table 4.3.1 below the hazard ratio calculated for the age group in the row compared to the age group in the column is displayed (95% confidence interval for the hazard ratio is attached as well).

The hazard of a young age group customer (age less than 26 years) cancelling the service is 2.31 times the hazard of a middle age group customer (age from 26 to 40 years) with 95% confidence limits (2.09, 2.55) and 3.81 times the hazard of an old age

Table 4.2.2: The joint p-value calculated for the grouping variables

The variable	The joint p-value
Age group	0.0001
Language	0.0001
Direct marketing city	0.003

Table 4.3.1: The hazard ratio calculated for the age group in the row compared to the age group in the column (95% confidence interval for the hazard ratio is attached as well).

		Age group		
		Less than 26 years	26 to 40 years	More than 40 years
Age Group	Less than 26 years	1	2.31 (2.09,2.55)	3.81 (3.43 , 4.23)
	26 to 40 years	0.43 (0.39,0.48)	1	1.65 (1.52 , 1.80)
	More than 40 years	0.26 (0.24,0.29)	0.61 (0.56,0.66)	1

group customer with 95% confidence limits (3.43, 4.23). The hazard of a middle age group customer (age from 26 to 40 years) cancelling the service is 0.43 times the hazard of a young age group customer with 95% confidence limits (0.39, 0.48) and is 1.65 times the hazard of an old age group customer with 95% confidence limits (1.52, 1.80). The hazard of an old age group customer (age more than 40 years) cancelling the service is 0.26 times the hazard of a young age group customer with 95% confidence limits (0.24, 0.29) and is 0.61 times the hazard of a middle age group customer with 95% confidence limits (0.56, 0.66). These results show that customers of age less than 26 years have the highest risk of cancelling the service when comparing them to customers from other age groups. The reason could be due to the lack of having a sustainable source of income or could be due to their enthusiasm to look for new possibilities. Either way, a careful plan has to be set in place in order to retain the young age group customers. The second highest risk of cancellation is found to be in the middle age group customers and the least risk of cancelling the service is for the old age group customers.

When looking at the confidence interval for the hazard ratio of the age group in the row compared to the age group in the column no overlaps are observed. This suggests that the age group (with its current intervals) is a strong differential factor that differentiates between the loyal customers and the disloyal customers. Age group, therefore, can be used effectively to segment the customers and design retention strategies.

Table 4.3.2: The hazard ratio calculated for language in the row compared to language in the column (95% confidence interval for the hazard ratio is attached as well).

		Language		
		English	Afrikaans	Other
Language	English	1	0.71 (0.60,0.83)	0.69 (0.17,2.77)
	Afrikaans	1.41 (1.20,1.67)	1	0.98 (0.24,3.95)
	Other	1.44 (0.36,5.78)	1.02 (0.25,4.13)	1

Table 4.3.2 presents the hazard ratio of language in the row compared to language in the column with a 95% confidence interval. This enables the comparison of risk of

cancelling the service across languages. The table shows that customers who stated English as their preferred service language had 0.71 times the risk of those who stated Afrikaans, with a 95% confidence interval (0.60, 0.83) and 0.69 times the risk of those who stated other languages with a 95% confidence interval (0.17, 2.77). Customers who stated Afrikaans as their language have 1.41 times the risk of those who stated English with a 95% confidence interval (1.20, 1.67), and 0.98 times the risk of those who stated other language, with a 95% a confidence interval (0.24, 3.95). A customer with other languages has 1.44 times the risk of a customer with English language, with a 95% confidence interval (0.36, 5.78) and 1.02 times the risk of a customer with Afrikaans language, with a 95% confidence interval (0.25, 4.13). However, overlaps in the confidence intervals are seen between English and other languages as well as Afrikaans and other languages; the confidence intervals are wide in both cases and include one. This could be explained by a small number of customers with other languages. Customers with English have been significantly different from Afrikaans customers in their risk of cancelling the service. In general, the language factor is still a significant factor in differentiating between customers with a high likelihood of leaving the service provider (with special emphasises on English and Afrikaans).

In table 4.3.3, the hazard ratio is presented for a customer subscribed to a service in the city in the row compared to a customer subscribed to a service in the city in the column (with a 95% confidence interval). This application tries to explain the spatial effect of marketing, that is, the impact of the marketing city. Customers subscribed to a service in the city of Witwatersrand and the city of Durban have shown the highest risk of service cancellation compared to all other cities. This suggest that the marketing city is an important factor to consider when we design marketing campaigns (retention program), but not as important as the age factor. The hazard of cancelling the service in Witwatersrand compared to Durban is 1.02 with a 95% confidence interval (0.88, 1.20). The city of Johannesburg has the smallest risk of service cancellation compared to all other cities.

Table 4.3.3: The hazard ratio calculated for marketing city in the row compared to marketing city in the column (95% confidence interval for the hazard ratio is attached as well).

		Marketing city					
		Cape Town	Durban	Johannesburg	Pretoria	Witwatersrand	Others
Marketing city	Cape Town	1	0.91 (0.78,1.06)	1.13 (1.00,1.29)	0.94 (0.82,1.09)	0.89 (0.78,1.00)	1.01 (0.90,1.13)
	Durban	1.10 (0.94, 1.29)	1	1.25 (1.07,1.47)	1.04 (0.88,1.24)	0.98 (0.84,1.14)	1.11 (0.96,1.29)
	Johannesburg	0.88 (0.78,1.00)	0.80 (0.68,0.94)	1	0.83 (0.72,0.96)	0.78 (0.69,0.88)	0.89 (0.79,1.00)
	Pretoria	1.06 (0.92, 1.21)	0.96 (0.81,1.13)	1.20 (1.04,1.38)	1	0.94 (0.82,1.07)	1.07 (0.94,1.21)
	Witwatersrand	1.13 (1.00, 1.28)	1.02 (0.88,1.20)	1.28 (1.13,1.45)	1.07 (0.93,1.23)	1	1.14 (1.02,1.28)
	Others	0.99 (0.88,1.11)	0.90 (0.77,1.04)	1.12 (1.00,1.26)	0.88 (0.82,1.06)	0.88 (0.78,0.98)	1

4.3.2 The use of the Nelson-Aalen integrated hazard to understand the risk of cancellation of service over time

In this sub-section, the cumulative hazard of service cancellation is calculated for demographic and usage-related variables using the nonparametric method of Nelson-Aalen. The Nelson-Aalen function is calculated over the full data and evaluated at indicated times (the function was evaluated at 1, 13, 25, 37, 49, 61, 73, 85 and 97 months). The reason for evaluating the hazard at these time points is to keep the presentation of the table simple and practical (by evaluating the hazard at the end of the first month and then after each year).

Table 4.3.4: The cumulative hazard of service cancellation over time by gender (95% confidence interval is attached).

Time (in months)	Gender	
	Male	Female
1	0.0047 (0.0038 , 0.0058)	0.0063 (0.0048 , 0.0083)
13	0.0416 (0.0385 , 0.0449)	0.0560 (0.0507 , 0.0618)
25	0.0775 (0.0731 , 0.0822)	0.1063 (0.0985 , 0.1147)
37	0.1377 (0.1314 , 0.1444)	0.1610 (0.1508 , 0.1720)
49	0.1967 (0.1883 , 0.2054)	0.2264 (0.2128 , 0.2408)
61	0.2563 (0.2453 , 0.2678)	0.2980 (0.2794 , 0.3179)
73	0.3273 (0.3124 , 0.3430)	0.3805 (0.3546 , 0.4082)
85	0.3927 (0.3709 , 0.4158)	0.4487 (0.4101 , 0.4910)
97	0.4209 (0.3905 , 0.4537)	0.5799 (0.3777 , 0.8903)

When gender is considered, the hazard of a female cancelling the service is significantly higher than the hazard of a male cancelling the service. This has been the case with exception to the first, 85th and 97th month. This result, together with the result from the Cox model, suggests that gender could be considered as a potential factor in predicting customer risk of cancelling the service. Therefore, to fight the cancellation, the marketing department should consider the gender of the customer as an important factor when designing retention strategies.

Table 4.3.5: The cumulative hazard of service cancellation over time by age group (95% confidence interval is attached).

Time (in months)	Age group		
	Less than 26 years	26-40 years	More than 40 years
1	0.0137 (0.0093 , 0.0201)	0.0063 (0.0047 , 0.0083)	0.0051 (0.0038 , 0.0070)
13	0.0928 (0.0791 , 0.1090)	0.0709 (0.0649 , 0.0776)	0.0434 (0.0388 , 0.0485)
25	0.2147 (0.1910 , 0.2413)	0.1339 (0.1248 , 0.1436)	0.0822 (0.0755 , 0.0895)
37	0.4545 (0.4142 , 0.4987)	0.2289 (0.2157 , 0.2428)	0.1352 (0.1260 , 0.1451)
49	0.7133 (0.6552 , 0.7767)	0.3157 (0.2989 , 0.3334)	0.1850 (0.1736 , 0.1972)
61	0.8096 (0.7421 , 0.8833)	0.3978 (0.3765 , 0.4203)	0.2405 (0.2258 , 0.2561)
73	0.9093 (0.8282 , 0.9983)	0.4846 (0.4569 , 0.5140)	0.3189 (0.2979 , 0.3414)
85	0.9474 (0.8567 , 1.0476)	0.5719 (0.5305 , 0.6164)	0.3855 (0.3533 , 0.4206)
97	1.0051 (0.8875 , 1.1382)	0.6313 (0.5614 , 0.7099)	0.4500 (0.3625 , 0.5586)

Table 4.3.5 shows the cumulative hazard of service cancellation over time for different age groups. The younger customers have shown a very high risk of cancellation. As the customer's age increases the chance of service cancellation decreases. The risk of churn is more than two times higher for the age group of a subscriber with an age less than 26 years than for the subscriber with age more than 40 years. No overlaps in the estimate of the cumulative hazard of the customer with an age less than 26 years and customers with age more than 40 years were observed. This suggests that the age group factor is a variable that can be used to understand service cancellation.

The Nelson-Aalen cumulative hazard shows that the risk of cancellation over time differs significantly between English and Afrikaans speaking customers. Customers with Afrikaans showed a higher risk than customers with English. No overlaps in the

confidence intervals of hazard of English and Afrikaans speakers are seen, except at the 97th month. The overlap at the 97th month could be due to a small number of customers in a later stage of the study, which resulted in a high standard error and, hence wider confidence limits. Using Nelson-Aalen for the hazard by language suggests that language can be used as a potential factor to predict service cancellation.

From table 4.3.7, it can be seen that when the WiFi usage factor is considered, customers who use WiFi showed a lower point estimate of the cumulative hazard than

Table 4.3.6: The cumulative hazard of service cancellation over time by language (95% confidence interval is attached).

Time (in months)	Language		
	English	Afrikaans	Others
1	0.0044 (0.0037, 0.0052)	0.0115 (0.0070 , 0.0191)	0
13	0.0389 (0.0365, 0.0414)	0.1127 (0.0941 , 0.1351)	0.0588 (0.0083 , 0.4176)
25	0.0747 (0.0712, 0.0783)	0.1818 (0.1554 , 0.2126)	0.2469 (0.0766 , 0.7955)
37	0.1252 (0.1205, 0.1302)	0.2767 (0.2382 , 0.3214)	0.2469 (0.0766 , 0.7955)
49	0.1770 (0.1709, 0.1833)	0.3382 (0.2871 , 0.3983)	0.2469 (0.0766 , 0.7955)
61	0.2401 (0.2320, 0.2485)	0.3748 (0.3064 , 0.4585)	
73	0.3128 (0.3021, 0.3239)	0.4452 (0.3479 , 0.5698)	
85	0.3911 (0.3721, 0.4111)	0.6264 (0.4680 , 0.8384)	
97	0.4373 (0.3980, 0.4804)	0.6264 (0.4680 , 0.8384)	

Table 4.3.7: The cumulative hazard of service cancellation over time by WiFi usage (95% confidence interval is attached).

Time (in months)	WiFi usage	
	No	Yes
1	0.0046 (0.0039 , 0.0055)	0.0052 (0.0007 , 0.0366)
13	0.0414 (0.0391 , 0.0440)	0.0325 (0.0146 , 0.0725)
25	0.0779 (0.0745 , 0.0815)	0.0456 (0.0227 , 0.0914)
37	0.1286 (0.1239 , 0.1335)	0.0690 (0.0379 , 0.1255)
49	0.1792 (0.1733 , 0.1854)	0.1050 (0.0625 , 0.1762)
61	0.2400 (0.2321 , 0.2481)	0.1261 (0.0773 , 0.2058)
73	0.3085 (0.2982 , 0.3192)	0.1953 (0.1251 , 0.3051)
85	0.3715 (0.3556 , 0.3881)	0.2317 (0.1501 , 0.3577)
97	0.4105 (0.3754 , 0.4488)	

customers who do not use WiFi. However, frequent overlaps are seen in the interval estimates of hazard for those who use WiFi and those who do not use WiFi.

Table 4.3.8: The cumulative hazard of service cancellation over time by IT background (95% confidence interval is attached).

Time (in months)	IT Background	
	No	Yes
1	0.0048 (0.0041 , 0.0057)	0
13	0.0429 (0.0405 , 0.0455)	0.0019 (0.0005 , 0.0076)
25	0.0808 (0.0772 , 0.0845)	0.0019 (0.0005 , 0.0076)
37	0.1329 (0.1281 , 0.1379)	0.0019 (0.0005 , 0.0076)
49	0.1836 (0.1775 , 0.1899)	
61	0.2441 (0.2362 , 0.2524)	
73	0.3128 (0.3024 , 0.3234)	
85	0.3759 (0.3600 , 0.3924)	
97	0.4146 (0.3796 , 0.4529)	

Table 4.3.8, shows that customers with an IT background have a lower risk of service cancellation than customers with no IT background. An IT background appears to be a significant variable in order to differentiate between customers who are likely to cancel their service and those who are more likely to remain loyal.

Table 4.3.9: The cumulative hazard of service cancellation over time by service usage purpose (95% confidence interval is attached).

Time (in months)	Service usage purpose (segment)	
	Private	Private & Business
1	0.0088 (0.0072 , 0.0107)	0.0007 (0.0004 , 0.0012)
13	0.0708 (0.0654 , 0.0767)	0.0135 (0.0119 , 0.0154)
25	0.1076 (0.1003 , 0.1154)	0.0405 (0.0374 , 0.0437)
37	0.1383 (0.1294 , 0.1478)	0.0860 (0.0814 , 0.0910)
49	0.1770 (0.1662 , 0.1886)	0.1286 (0.1224 , 0.1351)
61	0.2152 (0.2020 , 0.2293)	0.1598 (0.1522 , 0.1679)
73	0.2745 (0.2569 , 0.2933)	0.2170 (0.2066 , 0.2280)
85	0.3389 (0.3106 , 0.3698)	0.2762 (0.2588 , 0.2948)
97	0.4329 (0.3135 , 0.5977)	0.3003 (0.2759 , 0.3268)

Customers who use the service for both private and business purposes have less

hazard of service cancellation than those who use the service for only private purposes. Very few overlaps in the confidence interval of the cumulative hazard are seen. This suggests that the service usage purpose is important and a significant factor in order to differentiate between loyal and risky customers. Therefore, it can be used as a segmentation variable and a variable to consider when retention programs have to be designed.

4.4 Understanding customer survival probabilities using the Kaplan-Meier method

In the previous sections the risk of cancellation of the service was analysed using both the Cox model and Nelson-Aalen methods. In this section, customer loyalty and survival probabilities are analysed using the Kaplan-Meier method. The Kaplan-Meier analysis will answer the questions of who are likely to stay, and quantify the probability of survival for each subgroup of customers empirically. In addition to helping managers design strategies to retain customers, this analysis will help managers in planning customers' acquisition as well. The results of this section are presented graphically; p-values calculated from the log-rank test will be attached as well.

Figure 4.4.1 presents customers' survival curves by gender. Male customers have shown better survival chances than female customers. No overlaps are seen except at a very late stage of the analysis time. The p-value calculated from the log-rank test to test the equality of the survival curves is 0.0001. The gender variable is considered to be a significant factor in order to differentiate the loyal from the disloyal customers.

Figure 4.4.2 presents the customers' survival curve by age group. The young customers have less survival chance while the older customers seem to be more loyal. The age factor has shown a high potential in differentiating between the loyal and risky subscribers. No overlaps in the survival confidence limits were seen except after 93 months. The p-value calculated from the log-rank test to test the equality of the survival curves is 0.0001.

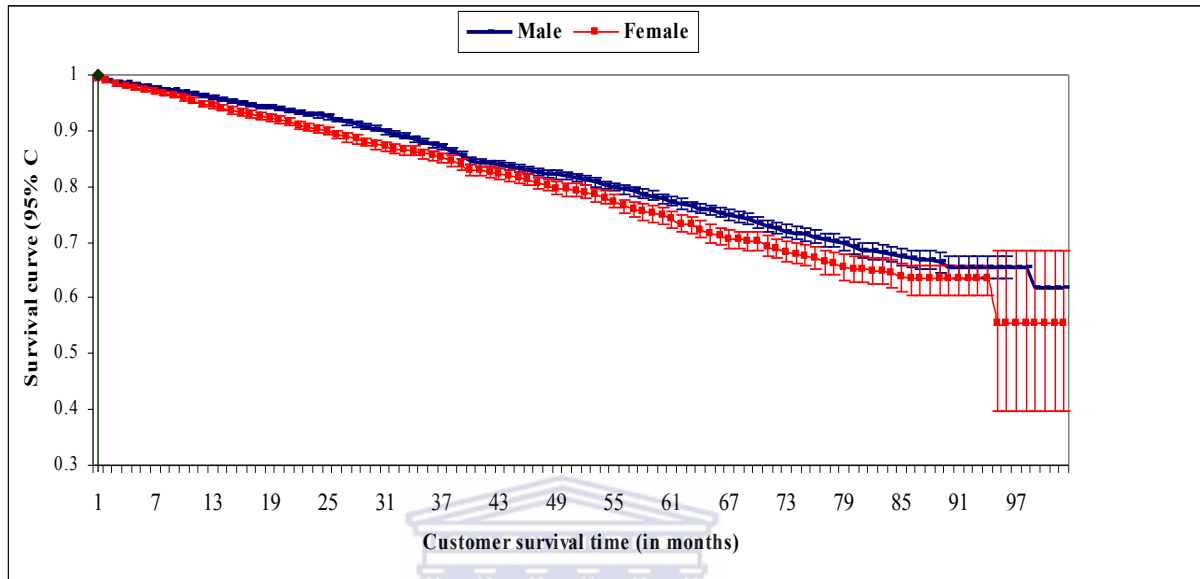


Figure 4.4.1: Customer survival probabilities by gender (95% confidence interval is attached)

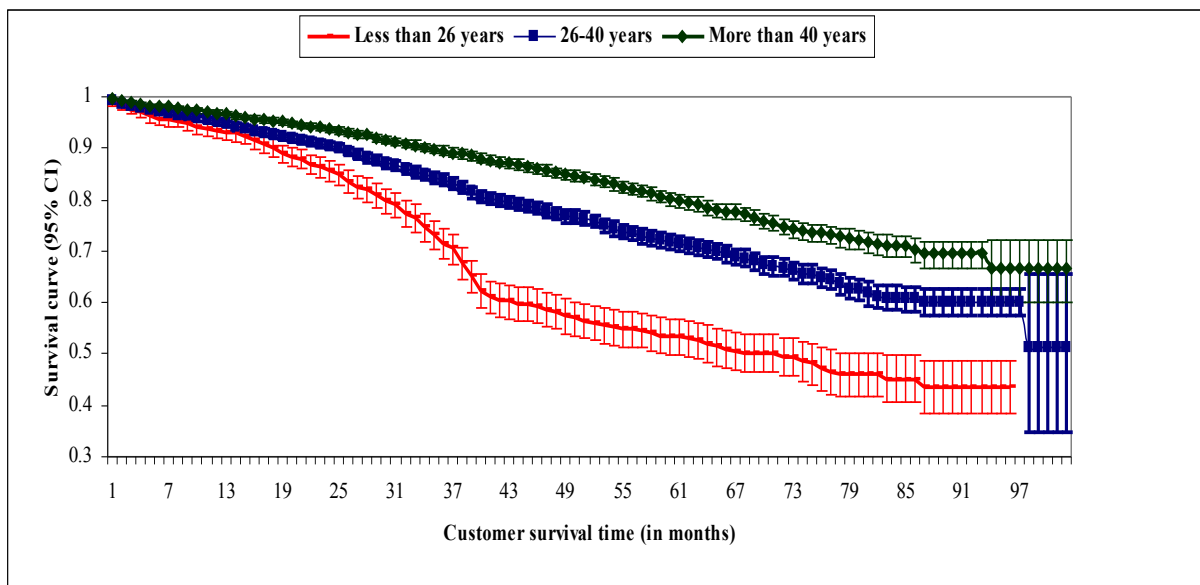


Figure 4.4.2: Customer survival probabilities by age group (95% confidence interval is attached)

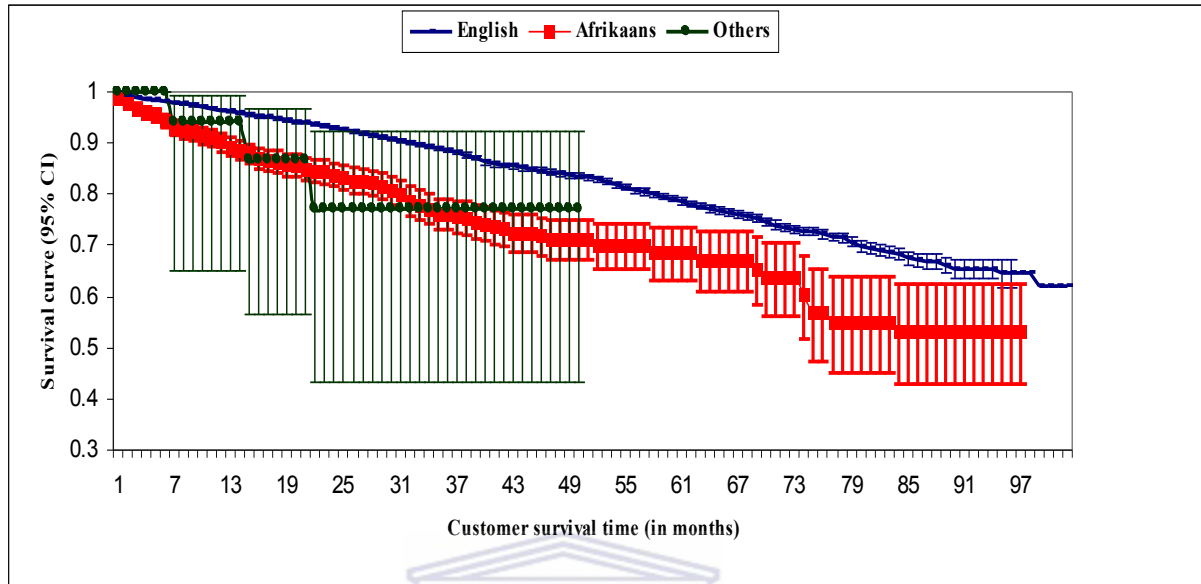


Figure 4.4.3: Customer survival probabilities by language (95% confidence interval is attached)

Figure 4.4.3 presents the customers' survival curve by language. Customers with Afrikaans language have less survival chance than those with English language. The language factor has shown potential in differentiating between risky and loyal subscribers; however, the survival curve for the category of "other language" has shown overlaps with the other two languages (English and Afrikaans). This could be due to the small number of customers in this category. The log-rank test for equality of the three survival curves gave a p-value equal to 0.0001.

In figure 4.4.4, the estimated survival curves for customers according to usage purposes are presented with 95% confidence limits. Customers who use the service for both private and business purposes have a better survival chance than those who use it only for private purposes. The survival curve confidence limits have shown a clear difference as there were no overlaps except at very few points towards the end of the study. The log-rank test for equality of the two survival curves gave a p-value equal to 0.0001.

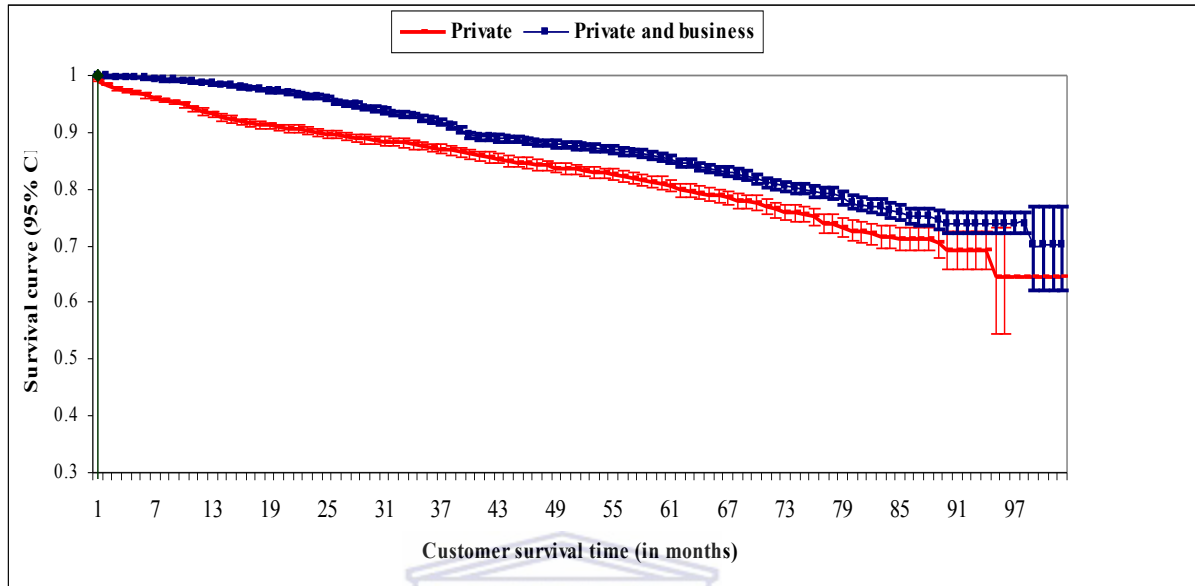


Figure 4.4.4: Customer survival probabilities according to usage purposes (95% confidence interval is attached)

When studying the interaction between variables included in the Cox model, only the interaction between gender and age group was found to be significant. Figure 4.4.5, 4.4.6 and 4.4.7 present the survival curve by gender for age group less than 26 years, 26 to 40 years and more than 40. The difference in survival curve of males and females in the age group of more than 40 years was statistically significant (unlike the other age groups where overlaps between males and females are observed).

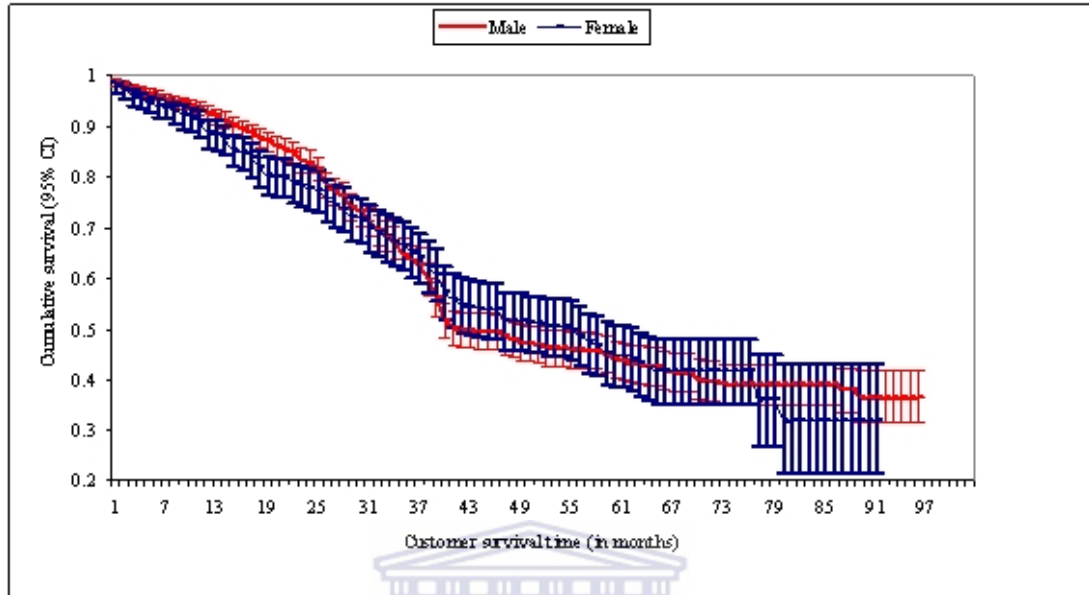


Figure 4.4.5: Customer survival probabilities by gender and in group less than 26 years (95% confidence interval is attached)

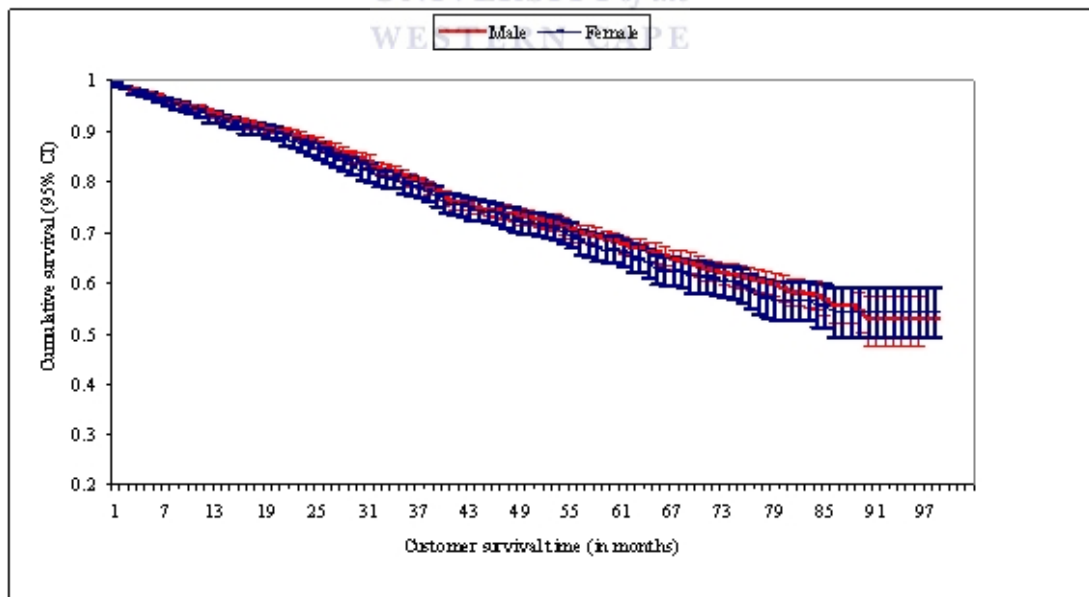


Figure 4.4.6: Customer survival probabilities by gender and in group 26 to 40 years (95% confidence interval is attached)

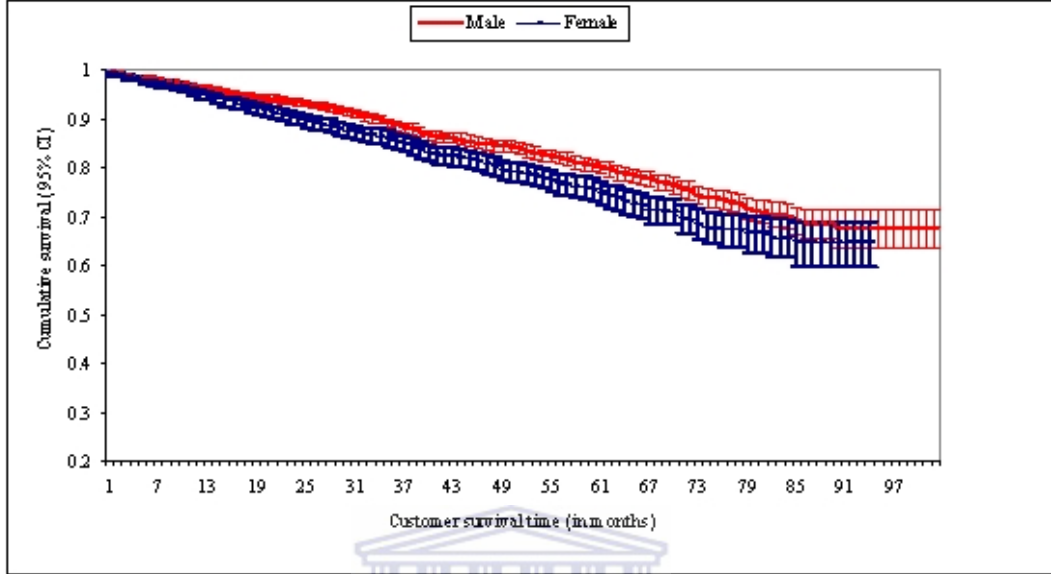


Figure 4.4.7: Customer survival probabilities by gender and in age group more than 40 years (95% confidence interval is attached)

4.5 Estimating the customer's mean survival time using both the non-parametric and parametric model

In this section, estimates of customer mean survival time are presented using both parametric and non-parametric methods. We estimate the customer mean survival time from parametric regression after fitting the parametric regression (without covariates) using the exponential, Weibull and gamma distributions. We estimate the mean from the Kaplan-Meier method as well. The estimate of the mean has been calculated for each level of the following covariates: “gender”, “age group” and “usage purpose”. In the parametric regression scenario, different models have been fitted for each level of the three mentioned covariates. The purpose of this exercise is, firstly, to find estimates of the mean survival time using various techniques. This will provide valuable information to managers to help understand their customer's lifetime and for comparing the mean survival time of their customers across the levels of each covariate. Secondly,

the results of this section will enable us to observe the variation in the estimates of the mean survival time using different methods of model fitting and approximation. This will help in understanding the adequacy and suitability of the current techniques in answering the question related to the estimation of the customer's mean survival time.

Table 4.5.1: Estimation of customer mean survival time (in months) by gender and type of the model

	Model used to calculate the customer mean survival time			
Gender	Kaplan-Meier	Exponential	Weibull	Gamma
Male	83.1 (82.4,83.8)	245.4 (236.4,254.8)	175.1 (166.4,184.2)	191.1 (182.2, 200.4)
Female	80.1 (78.7,81.6)	209.6 (198.7, 221.1)	171.2 (158.6, 184.8)	180.9 (168.3, 194.2)

Table 4.5.2: The range of the point estimate of customer mean survival time for different modelling approaches and by gender

Gender	Parametric models	Nonparametric and parametric models
Male	70.3	162.3
Female	38.4	129.5

Table 4.5.1 presents customer mean survival time by gender and types of the models used to estimate the mean. The non-parametric estimate of the mean survival time using the empirical data, is equal to 83.1 months with a 95% confidence interval (82.4, 83.8) for male customers, and 80.1 months, with a 95% confidence interval (78.7, 81.6) for females. The estimate of the male mean survival time was 245.4, 175.1 and 191.1 months assuming exponential, Weibull and gamma distributions respectively (a 95% confidence interval is attached). The estimate of the female mean survival time was 209.6, 171.2 and 180.9 months assuming exponential, Weibull and gamma distributions respectively (a 95% confidence interval is attached). The range of the estimates of customer mean survival time assuming exponential, Weibull and gamma distributions was 70.3 months for males and 38.4 months for females. The range of the estimates of the customer mean survival from nonparametric and parametric methods was 162.3 months for males and 129.5 for females. Apart from the difference between male and female mean survival time, this table showed a large variation in the estimate of the

mean survival time when different parametric methods are assumed. It also shows a large difference between the Kaplan-Meier estimate and the estimates obtained from the parametric distributions.

Table 4.5.3: Estimation of customer mean survival time by age group and type of the model

	Model used to calculate the customer mean survival time			
Age group	Kaplan-Meier	Exponential	Weibull	Gamma
< 26 years	58.3 (56.1,60.4)	83.4 (77.3,89.7)	67.2 (63.0,71.7)	70.1 (65.6,75.0)
26-40 years	74.2 (73.0,75.3)	156.4 (149.2, 164.1)	127.8 (120.7,135.3)	134.2 (127.1, 141.8)
> 40 years	83.5 (82.5,84.6)	252.0 (238.5,266.3)	190.0 (175.8, 205.4)	205.0 (190.7,220.3)

Table 4.5.4: The range of the point estimate of customer mean survival time for different modelling approaches and by age group

	The range of estimate of customer mean survival time	
Age group	Parametric models	Nonparametric and parametric models
Less than 26 years	16.2	25.1
26-40 years	28.6	82.2
More than 40 years	62	168.5

Table 4.5.3 presents customer mean survival time by age group and types of the model used to estimate the mean survival time. The Kaplan-Meier estimate of the mean survival time, using empirical data, is equal to 58.3 months with a 95% confidence interval (56.1, 60.4) for customers of age less than 26 years, 74.2 months with a 95% confidence interval (73.0, 75.3) for customers with ages from 26 to 40 years and 83.5 months with a 95% confidence interval (82.5, 84.6) for customers with more than 40 years of age. The estimates of the mean survival time for a customer with age less than 26 years were 83.4, 67.2 and 70.1 months assuming exponential, Weibull and gamma distributions respectively (a 95% confidence interval is attached). The estimates of mean survival time for a customer in the age range from 26 to 40 years were 156.4, 127.8 and 134.2 months assuming exponential, Weibull and gamma distributions respectively

(a 95% confidence interval is attached). The estimates of mean survival time for a customer more than 40 years of age were 252.0, 190.0 and 205.0 months assuming exponential, Weibull and gamma distributions respectively (a 95% confidence interval is reported). The ranges of the estimates of mean survival time assuming exponential, Weibull and gamma distributions were 16.2, 28.6, and 62 months for age group less than 26 years, 26 to 40 years and more than 40 years respectively. The ranges of the estimates of mean survival time from nonparametric and parametric methods were 25.1, 82.2, and 168.5 months for age group less than 26 years, 26 to 40 years and more than 40 years respectively. Apart from the differences in mean survival time across age groups, this result showed a large variation in the estimate of the mean survival time when different parametric methods were used. It also shows a large difference between the Kaplan-Meier estimate and the estimates from the parametric distributions.

Table 4.5.5: Estimation of customer mean survival time by usage purpose and type of the model

	Model used to calculate the customer mean survival time			
Usage purpose	Kaplan-Meier	Exponential	Weibull	Gamma
Private	84.1 (83.0,85.2)	246.4 (233.5,260.0)	337.2 (301.8,376.8)	309.1 (281.2, 339.7)
Private/Business	89.0 (88.4,89.6)	372.9 (357.5, 388.9)	171.0 (163.0, 179.5)	196.5 (187.4, 206.0)

Table 4.5.6: The range of the point estimate of customer mean survival time for different modelling approaches and by usage purpose

	The range of estimate of customer mean survival time	
Usage purpose	Parametric models	Nonparametric and parametric models
Private	90.8	253.1
Private/Business	201.9	283.9

Table 4.5.5 presents customer mean survival time by usage purpose and the types of model used to estimate the mean. The non-parametric estimate of the mean survival time, using the empirical data, is equal to 84.1 months with a 95% confidence interval

(83.0, 85.2) for a customer who uses the service for private purposes and 89.0 months with a 95% confidence interval (88.4, 89.6) for one who uses the service for both private and business purposes. The estimates of the private purposes customer's mean survival time were 246.4, 337.2 and 309.1 months assuming exponential, Weibull and gamma distributions respectively (a 95% confidence interval is attached). For the customer who uses the service for both business and private purposes, the estimates of the mean survival time were 372.9, 171.0 and 196.5 months assuming exponential, Weibull and gamma distributions respectively (a 95% confidence interval is attached). The range of the estimates of the mean survival time assuming exponential, Weibull and gamma distributions was 90.8 months for customers who use the service for private purposes and 201.9 months for customers who use the service for both business and private purposes. The ranges of the estimates of mean survival time from nonparametric and parametric methods were 253.1 months for customers who use the service for private purposes and 283.9 months for customers who use the service for both business and private purposes. Apart from the difference between customers' mean survival time due to their usage purposes, this result showed a large variation in the estimate of the mean survival time when different parametric methods are assumed. It also shows a large difference between the Kaplan-Meier estimate and the estimates obtained from the parametric distributions.

4.6 Summary and discussion

This chapter aimed at understanding the business and the methodological insights in the process of applying survival analysis techniques to analyse customer survival time in subscription-based businesses using a data set that contains demographic and usage related variables. The results showed the importance of demographic and usage-related factors in understanding customer survival time and customer loyalty. The stratified Cox model (stratification variable was a customer usage purpose segment) is used to identify the significant variables that affect the customer survival time.

The results showed that gender, language, age, marketing city and WiFi usage are statistically significant variables in predicting customer risk of cancellation at a 0.05 significance level (the p-value associated with the IT background variable was equal to 0.076). The hazard ratios obtained from a stratified Cox model, together with the cumulative hazard obtained from the Nelson-Aalen method, were used to study and understand the customer's risk of cancellation of the service. With respect to gender, the risk of a female cancelling the service was found to be higher than the risk of a male cancelling the service. It is important for the marketing department to conduct further studies to check on gender-based preferences and expectations. The service provided to the customer can be personalised based on gender and a segment-based approach is necessary here.

Age appeared to be the most important factor that differentiates between the risky customers and the loyal customers. Big differences between the age groups in the hazard ratio, the cumulative hazard, and the survival probability were observed. Customers in the young age group have shown the highest risk of service cancellation; they have the highest hazard ratio in the stratified Cox analysis and the highest cumulative hazard in the Nelson-Aalen analysis. The Kaplan-Meier analysis of age group showed that the customers in the young age group have the lowest survival probabilities, followed by the middle age group, while the old age group customers have the highest survival chances. Whatever the explanation, a careful plan has to be set in place in order to retain the young age group customers. Customers with English appear to have better survival chances than those with Afrikaans or other languages. The significance of the marketing city in explaining the risk of cancellation of the service, can be due to the ability of this variable in explaining the relative importance of service to the customer and the level of socio-economic status of customers. The WiFi usage, IT background, and the usage purpose segment appeared to be significant, because they are highly connected to relative importance of the service to the customer, the ability of the customer to use the service and his/her level of service usage. The customer mean survival time estimated for each sub-population (the population has been sub-

divided into categories by demographic and usage-related variables) can be used to estimate the customer lifetime value for a customer that falls into that sub-population. The ultimate outcome of this process of studying customer survival time will be the understanding of the dynamic and the behaviour of customers with respect to their risk of cancellation survival probability and lifetime value. The analysis has motivated the use of the concept of market segmentation. Market segmentation will be based on a demographic and usage-related customer profile. Although the concept of market segmentation is well understood and studied in various fields of business, not many studies have been done in the particular area of research investigated here. However, the motivation behind market segmentation in this setting will stay the same as in other settings in the current literature and that is to have a better understanding of customers in order to personalise products and have customer specific marketing strategies (Wind, 1978; Badgett and Stone, 2004; Weinstein, 2004; Gopalan, 2007). The relevance and importance of the demographic and usage-related variables in understanding customers (segmenting customers according to their characteristics) that come out of these studies are in agreement with other research done in similar studies (Weinstein, 2002; Bruwer and Elton, 2007; Rugimbana, 2007; Encinas *et al.*, 2007), but it is the first of its kind in the literature on survival analysis in subscription-based businesses.

The results on the estimates for customer mean survival time were obtained via different non-parametric and parametric approaches; namely, the Kaplan-Meier method and the exponential, Weibull and gamma regression models. The estimates of the means from exponential regression, Weibull regression, and gamma regression vary greatly. The assumption imposed on the distribution of the survival time is very important. This suggests that a careful investigation of the method of extrapolating the survival curve beyond the empirical data and of the choice of the parametric models is extremely important especially in this type of business problems where a high degree of censoring is expected. In addition to the differences in customer mean survival time that are due to the use of different methodologies, results showed the difference in

the estimates that were due to different demographic and usage-related profiles of the customers.

In conclusion, we suggest that in estimating the customer mean survival time one would prefer nonparametric methods with a careful plan for dealing with extrapolation issues. However, from our understanding of the fast dynamics of the market and the fast change in customer behaviour, a conditional mean survival time based on the empirical distribution of the data - for each sub-population and for a reasonable time horizon - will be the way forward. This will enable us to get accurate and practically useful inferences. Therefore, we dedicate a crucial part of research to the issue of extrapolating the survival curve (see chapter five for this contribution).



Chapter 5

Extrapolation of the survival curve

This chapter is focused on the extrapolation of the survival curve beyond the last observed failure time. The chapter is composed of seven main sections. The motivation behind this exercise is presented in section one. In section two, the proposed function is stated and its mathematical accuracy is checked in section three. In section four, we derive the standard error of the estimate of the proposed function. Expression for the conditional survival function and conditional mean survival time are given in section five. An application on a real data set is presented in section six and we summarise the chapter in section seven.

5.1 Motivation

The practical motivation for extrapolating the survival curve beyond the empirical distribution originates from two issues, that of calculating survival probabilities beyond the empirical data and of calculating the conditional mean survival time at a specific point in time and for a specific time window in the future. These two issues are of importance in the business environment, because the survival probability and the mean survival time are the main components used in calculating customer lifetime value and, hence, customer equity.

From a methodological perspective, different methods of extrapolation might give very different forecasting figures and this has been found clearly in chapter four. In most situations, we might not have a clear understanding of the dynamic of the problem at hand, especially in a business setting where the customers' behaviour and the market characteristics change fairly fast. Therefore, we favour the survival probability estimates obtained from the Kaplan-Meier method and use them as the basis for extrapolation.

5.2 The proposed extrapolation function

Denote the customer survival time by t where $t \geq 0$; t is measured in months. Denote the last observed failure time by τ . The Kaplan-Meier estimate of the survival curve over the interval $0 \leq t \leq \tau$ is

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i}; \quad 0 \leq t \leq \tau \quad (5.2.1)$$

(Kaplan and Meier, 1958). The motivation behind the use of the Kaplan-Meier method is to avoid incorrect assumptions about the underlying distribution of the customer survival time. But this method does not estimate the survival curve beyond the last failure time. Therefore, it underestimates the mean survival time if the last observed time point is censored time. The problem becomes worse when an extreme right censoring exists which is the case in most business data (it was 85% for the data used in chapter four). Our aim is to find a suitable extrapolation function to estimate the survival probability beyond the last failure time τ .

The choice of the proposed extrapolation function was made in a way that covers most of the common scenarios of risk of service cancellation; that is increasing risk of cancellation over time, decreasing risk of cancellation over time and constant risk of cancellation over time. To maximise the precision we made the estimate of the proposed extrapolation function and the Kaplan-Meier to be equal at $t = \tau$. The

proposed function is, therefore:

$$S(t) = S(\tau)e^{\beta(\tau-\tau_0)^\alpha - \beta(t-\tau_0)^\alpha}; \quad t > \tau \quad (5.2.2)$$

Estimates of the parameters of the extrapolation function α and β are $\hat{\alpha}$ and $\hat{\beta}$. They are obtained by minimising the following objective function:

$$\sum_{i=1}^n \left\{ \ln \left[\hat{S}(t_i) / \hat{S}(\tau) \right] - \beta [(\tau - \tau_0)^\alpha - (t_i - \tau_0)^\alpha] \right\}^2 \quad (5.2.3)$$

where $\tau_0 < \tau$ and $t_i \in [\tau_0, \tau]$. The data points t_i , $i = 1, 2, \dots, n$ are equally spaced. Sequential quadratic programming techniques can be used to solve the above minimisation problem for α and β (Han, 1976; Boggs *et al.*, 1982; Bonnons *et al.*, 1992).

Then, for a certain client in a certain sub-group, the estimate of the survivor function is

$$\hat{S}(t) = \begin{cases} \prod_{t(i) \leq t} \frac{n_i - d_i}{n_i}; & 0 \leq t \leq \tau \\ \hat{S}(\tau)e^{\hat{\beta}(\tau-\tau_0)^\alpha - \hat{\beta}(t-\tau_0)^\alpha}; & t > \tau \end{cases} \quad (5.2.4)$$

5.3 The mathematical check of the proposed survivor function

In this section, the accuracy of the proposed extrapolated survivor function will be checked mathematically on several kinds of lifetime data distributions. The distributions included in our investigations were the exponential, Weibull, gamma, mixture of two exponentials, mixture of two Weibulls, mixture of two gammas, mixture of the exponential and Weibull, mixture of exponential and gamma, and the mixture of Weibull and gamma distributions. These distributions represent a considerable variety of models that may reasonably be expected to fit the lifetime data.

The choice of the parameters of these distributions in each scenario is determined by how reasonable the mean customer survival time is that they give and how practical that would be in a business setting. The time points used to estimate the parameters of the extrapolation functions are $t = 90, 95, 100, 105, 110, 115$, and 120 ; with $\tau_0 = 90$ and $\tau = 120$. The choice of these data points were based on our observation of the real data set that we have studied in chapter four. These data points are reasonable time points that a customer could survive (time is measured in months).

5.3.1 Checking accuracy of the extrapolation function

The investigation is made by comparing the theoretical true survival probabilities with fitted survival probabilities. The fitted survival probabilities are the ones obtained from the extrapolation of the survivor function. The mathematical accuracy of the proposed function was tested over the time interval $[\tau_0, \tau + 60]$.

The method used to calculate the mathematical error is the maximum norm. The maximum norm is calculated for the difference between the theoretical survivor function and the fitted one. That is to find the $\max \{\|\delta S_t\|\}$, where $\delta S_t = S_{theo}(t) - S_{fit}(t)$, $S_{theo}(t)$ is the survival probabilities calculated from the theoretical distributions and $S_{fit}(t)$ is the survival probabilities calculated from the extrapolation function after estimating the parameters. The results of the mathematical accuracy check are presented as graphs and tables.

During the process of investigating the suitability of the proposed extrapolation function, SPSS 14.0 (SPSS, 2005), SPSS 15.0 (SPSS, 2006), Stata 8.0 (Sata, 2003) and wxMaxima 0.7.2 (open source) were used. Some of the graphs were produced using Microsoft Excel.

5.3.2 Results of the mathematical check

Results of the investigation of the mathematical accuracy of the proposed extrapolated survivor function are presented in this section. It is obvious that the extrapolation func-

Table 5.3.1: The theoretical and fitted survival probabilities for the gamma distribution

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.463	0.465	0.002
95	0.434	0.436	0.002
100	0.406	0.408	0.002
105	0.380	0.381	0.001
110	0.355	0.355	0.000
115	0.331	0.331	0.000
120	0.308	0.308	0.000
125	0.287	0.287	0.000
130	0.267	0.267	0.000
135	0.249	0.248	0.001
140	0.231	0.231	0.000
145	0.215	0.214	0.001
150	0.199	0.199	0.000
155	0.185	0.185	0.000
160	0.171	0.172	0.001
165	0.159	0.160	0.001
170	0.147	0.148	0.001
175	0.136	0.138	0.002
180	0.126	0.128	0.002

tion would fit perfectly the exponential survival function as the exponential survival is a special case of our extrapolation function; the same holds for the Weibull survival function. Therefore, we consider other survival functions, including the mixture distributions.

Gamma distribution

The true theoretical survival probability here follows a gamma survival function with a scale parameter equal to 0.02 and a shape parameter equal to 2. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 1.040$ and $\hat{\beta} = 0.012$. The theoretical and the fitted probabilities are given in table 5.3.1 and displayed in figure 5.3.1.

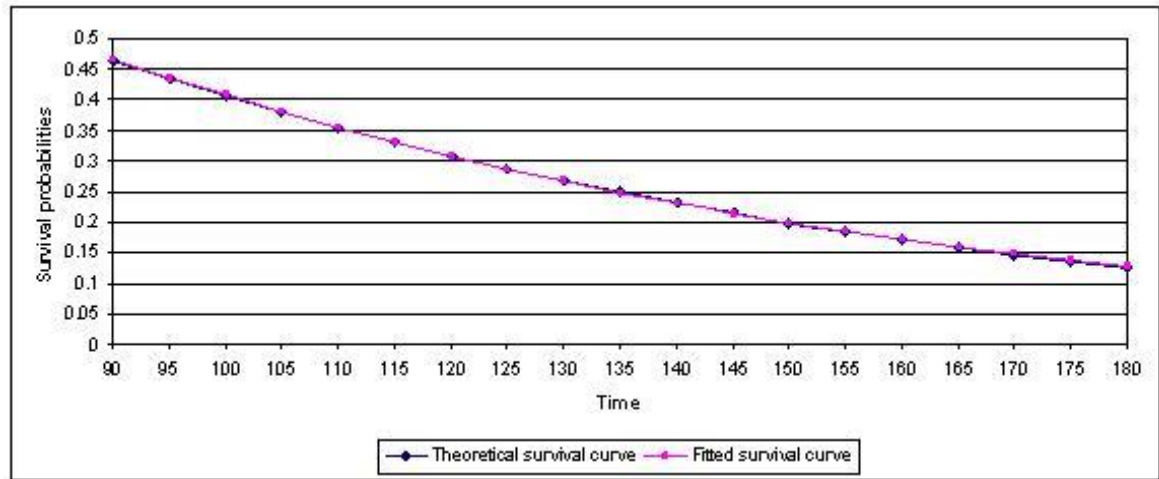


Figure 5.3.1: The theoretical and fitted survival curve for the gamma distribution

Mixture of two exponential distributions (1)

The true theoretical survival probability here follows a mixture of two exponential survival functions; that is 0.9 of an exponential survival with a scale parameter equal to 0.01 and 0.1 of an exponential survival function with a scale parameter of 0.02. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 0.999$ and $\hat{\beta} = 0.01$. The theoretical and the fitted probabilities are given in table 5.3.2 and displayed in figure 5.3.2.

Table 5.3.2: The theoretical and fitted survival probabilities for the mixture of two exponential distributions (1)

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.383	0.378	0.005
95	0.363	0.359	0.004
100	0.345	0.342	0.003
105	0.327	0.325	0.002
110	0.311	0.309	0.002
115	0.295	0.294	0.001
120	0.280	0.280	0.000
125	0.266	0.266	0.000
130	0.253	0.253	0.001
135	0.240	0.241	0.001
140	0.228	0.229	0.001
145	0.217	0.218	0.001
150	0.206	0.208	0.002
155	0.196	0.198	0.002
160	0.186	0.188	0.002
165	0.177	0.179	0.002
170	0.168	0.170	0.002
175	0.159	0.162	0.003
180	0.152	0.154	0.002

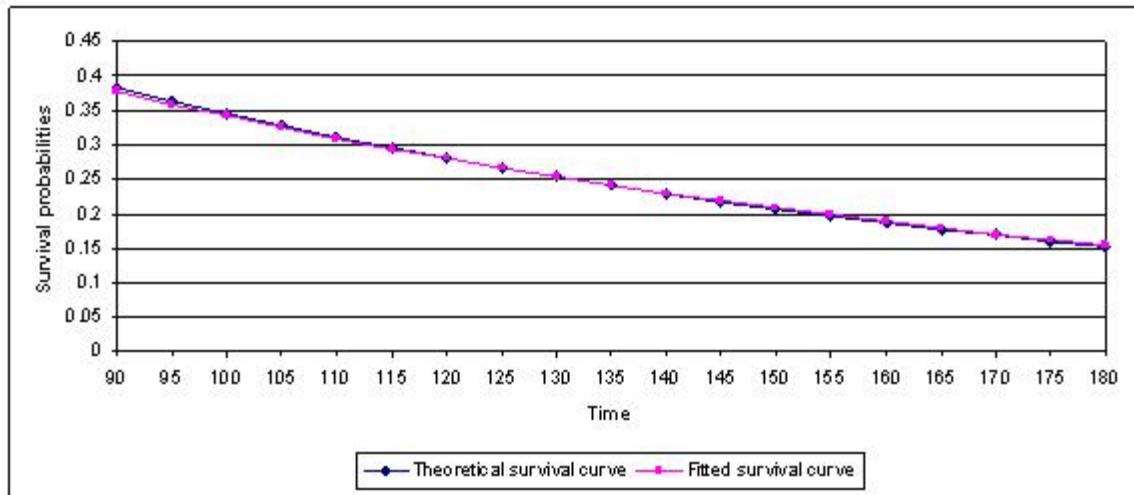


Figure 5.3.2: The theoretical and fitted survival curve for the mixture of two exponential distributions (1)

Mixture of two exponential distributions (2)

The true theoretical survival probability here follows a mixture of two exponential survival functions; that is 0.5 of an exponential survival with a scale parameter equal to 0.01 and 0.5 of an exponential survival function with a scale parameter of 0.02. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 0.983$ and $\hat{\beta} = 0.013$. The theoretical and the fitted probabilities are given in table 5.3.3 and displayed in figure 5.3.3.

Table 5.3.3: The theoretical and fitted survival probabilities for the mixture of two exponential distributions (2)

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.286	0.283	0.003
95	0.268	0.266	0.002
100	0.252	0.250	0.002
105	0.236	0.235	0.001
110	0.222	0.221	0.001
115	0.208	0.208	0.000
120	0.196	0.196	0.000
125	0.184	0.185	0.001
130	0.173	0.174	0.001
135	0.163	0.164	0.001
140	0.154	0.154	0.000
145	0.145	0.145	0.000
150	0.136	0.137	0.001
155	0.129	0.129	0.000
160	0.121	0.121	0.000
165	0.114	0.114	0.000
170	0.108	0.108	0.000
175	0.102	0.102	0.000
180	0.096	0.096	0.001

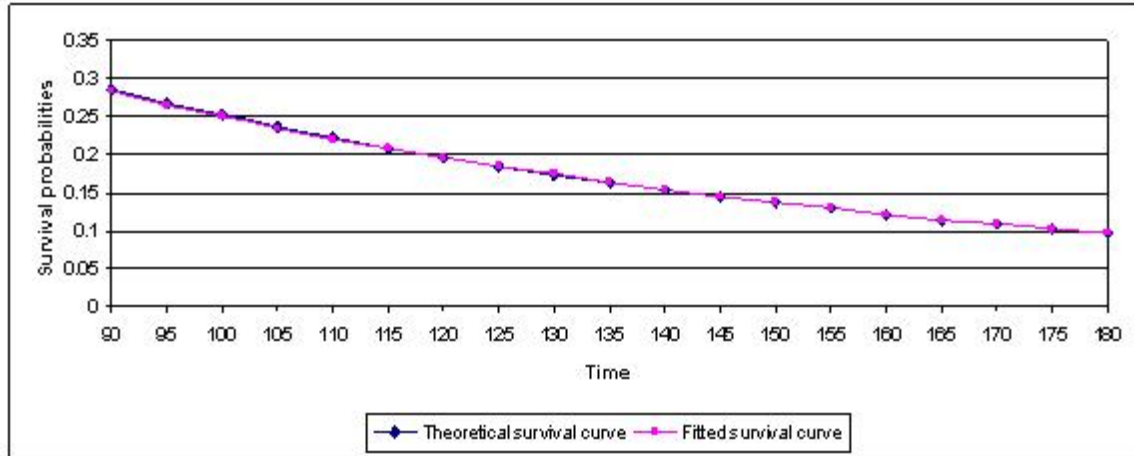


Figure 5.3.3: The theoretical and fitted survival curve for the mixture of two exponential distributions

Mixture of two Weibull distributions (1)

The true theoretical survival probability here follows a mixture of two Weibull survival functions; that is, 0.9 of a Weibull survival with a scale parameter equal to 0.005 and a shape parameter equal to 1.2 and 0.1 of a Weibull survival function with a scale parameter 0.005 and a shape parameter equal to 1.5. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 1.009$ and $\hat{\beta} = 0.015$. The theoretical and the fitted survival probabilities are given in table 5.3.4 and displayed in figure 5.3.4.

Table 5.3.4: The theoretical and fitted survival probabilities for the mixture of two Weibull distributions (1)

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.299	0.301	0.002
95	0.277	0.279	0.002
100	0.257	0.258	0.001
105	0.238	0.239	0.001
110	0.220	0.221	0.001
115	0.204	0.204	0.000
120	0.189	0.189	0.000
125	0.174	0.175	0.001
130	0.161	0.162	0.001
135	0.149	0.149	0.000
140	0.137	0.138	0.001
145	0.127	0.128	0.001
150	0.117	0.118	0.001
155	0.107	0.109	0.002
160	0.099	0.101	0.002
165	0.091	0.093	0.002
170	0.084	0.086	0.002
175	0.077	0.080	0.003
180	0.071	0.074	0.003

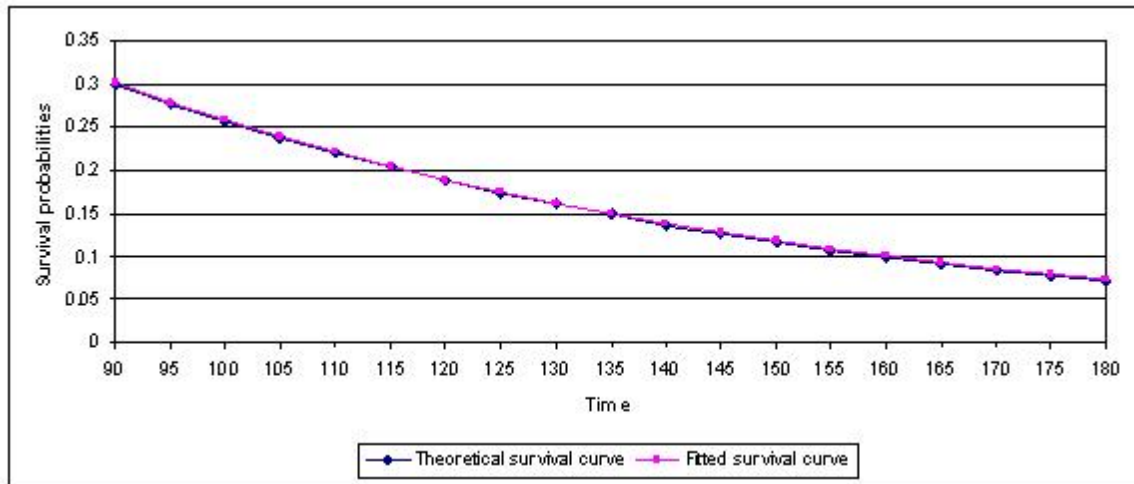


Figure 5.3.4: The theoretical and fitted survival curve for the mixture of two Weibull distributions (1)

Mixture of two Weibull distributions (2)

The true theoretical survival probability here follows a mixture of two Weibull survival functions; that is, 0.5 of a Weibull survival with a scale parameter equal to 0.01 and a shape parameter equal to 1.2 and 0.5 of a Weibull survival function with a scale parameter 0.008 and a shape parameter equal to 1.5. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 1.041$ and $\hat{\beta} = 0.010$. The theoretical and fitted survival probabilities are given in table 5.3.5 and displayed in figure 5.3.5.

Table 5.3.5: The theoretical and fitted survival probabilities for the mixture of two Weibull distributions (2)

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.479	0.479	0.000
95	0.453	0.454	0.001
100	0.428	0.429	0.001
105	0.405	0.405	0.000
110	0.382	0.382	0.000
115	0.360	0.360	0.000
120	0.339	0.339	0.000
125	0.319	0.319	0.000
130	0.300	0.301	0.001
135	0.282	0.283	0.001
140	0.265	0.266	0.001
145	0.248	0.250	0.002
150	0.233	0.235	0.002
155	0.218	0.221	0.003
160	0.204	0.208	0.004
165	0.190	0.196	0.006
170	0.178	0.184	0.006
175	0.166	0.173	0.007
180	0.155	0.162	0.007

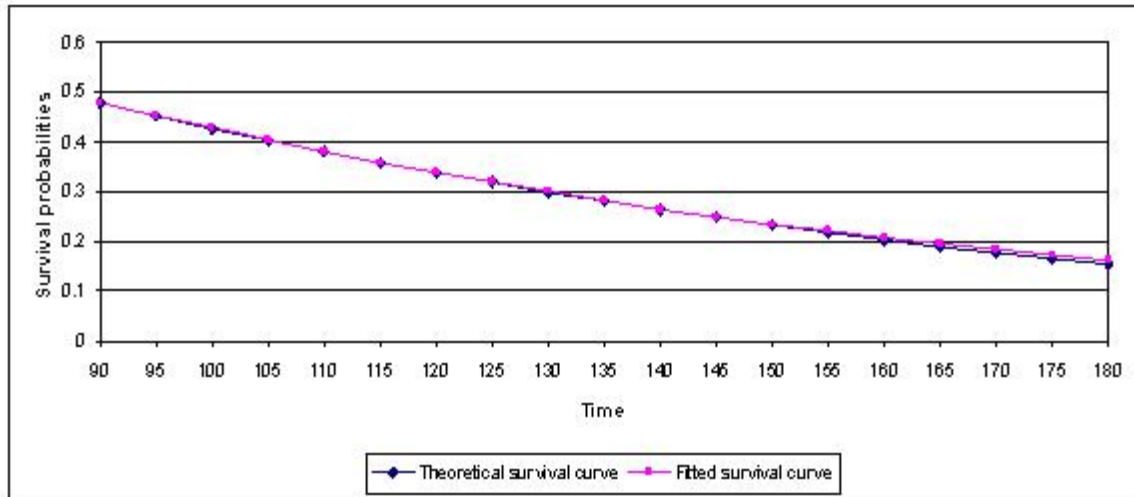


Figure 5.3.5: The theoretical and fitted survival curve for the mixture of two Weibull distributions (2)

Mixture of two gamma distributions

The true theoretical survival probability here follows a mixture of two gamma survival functions; that is 0.5 of a gamma survival with a scale parameter equal 0.02 and a shape parameter equal to 2 and 0.5 of a gamma survival function with a scale parameter 0.03 and a shape parameter equal to 3. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 1.041$ and $\hat{\beta} = 0.013$. The theoretical and fitted survival probabilities are given in table 5.3.6 and displayed in figure 5.3.6.

Table 5.3.6: The theoretical and fitted survival probabilities for the mixture of two gamma distributions

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.478	0.479	0.001
95	0.446	0.447	0.001
100	0.415	0.415	0.000
105	0.385	0.385	0.000
110	0.357	0.357	0.000
115	0.331	0.331	0.000
120	0.306	0.306	0.000
125	0.282	0.283	0.001
130	0.260	0.262	0.002
135	0.240	0.242	0.002
140	0.221	0.223	0.002
145	0.203	0.206	0.003
150	0.186	0.190	0.004
155	0.171	0.176	0.005
160	0.157	0.162	0.005
165	0.144	0.150	0.006
170	0.132	0.138	0.004
175	0.121	0.127	0.006
180	0.110	0.117	0.007

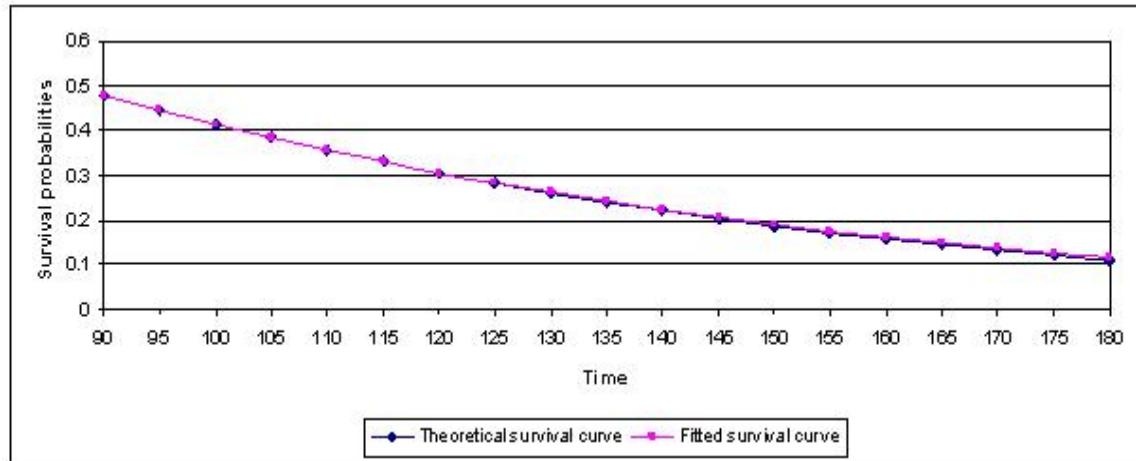


Figure 5.3.6: The theoretical and fitted survival curve for the mixture of two gamma distributions

Mixture of an exponential distribution and a Weibull distribution

The true theoretical survival probability here follows a mixture of an exponential survival function and a Weibull survival function; that is 0.5 of an exponential survival with a scale parameter equal to 0.01 and 0.5 of a Weibull survival function with a scale parameter 0.01 and a shape parameter of 1.5. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 0.983$ and $\hat{\beta} = 0.011$. The theoretical and fitted survival probabilities are given in table 5.3.7 and displayed in figure 5.3.7.

Table 5.3.7: The theoretical and fitted survival probabilities for the mixture of exponential and Weibull distributions

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.203	0.206	0.003
95	0.193	0.195	0.002
100	0.184	0.185	0.001
105	0.175	0.176	0.001
110	0.166	0.167	0.001
115	0.158	0.159	0.001
120	0.151	0.151	0.000
125	0.143	0.143	0.000
130	0.136	0.136	0.000
135	0.130	0.130	0.000
140	0.123	0.123	0.000
145	0.117	0.117	0.000
150	0.112	0.111	0.001
155	0.106	0.106	0.000
160	0.101	0.101	0.000
165	0.096	0.096	0.000
170	0.091	0.091	0.000
175	0.087	0.087	0.000
180	0.083	0.082	0.001

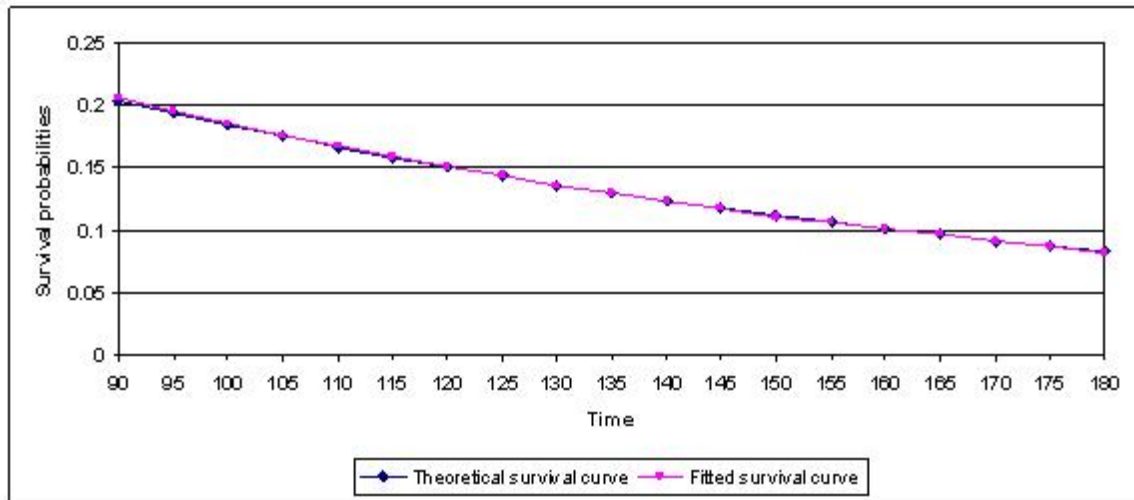


Figure 5.3.7: The theoretical and fitted survival curve for the mixture of exponential and Weibull distributions

Mixture of an exponential distribution and a gamma distribution

The true theoretical survival probability here follows a mixture of an exponential survival function and a gamma survival function; that is 0.5 of an exponential survival with a scale parameter equal to 0.01 and 0.5 of a gamma survival function with a scale parameter 0.02 and a shape parameter of 2. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 1.017$ and $\hat{\beta} = 0.011$. The theoretical and fitted survival probabilities are given in table 5.3.8 and displayed in figure 5.3.8.

Table 5.3.8: The theoretical and fitted survival probabilities for the mixture of exponential and gamma distributions

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.435	0.432	0.003
95	0.410	0.409	0.001
100	0.387	0.386	0.001
105	0.365	0.364	0.001
110	0.344	0.343	0.001
115	0.324	0.323	0.001
120	0.305	0.305	0.000
125	0.287	0.287	0.000
130	0.270	0.271	0.001
135	0.254	0.255	0.001
140	0.239	0.240	0.001
145	0.225	0.226	0.001
150	0.211	0.213	0.002
155	0.198	0.201	0.003
160	0.187	0.189	0.002
165	0.175	0.178	0.003
170	0.165	0.168	0.003
175	0.155	0.158	0.003
180	0.146	0.149	0.003

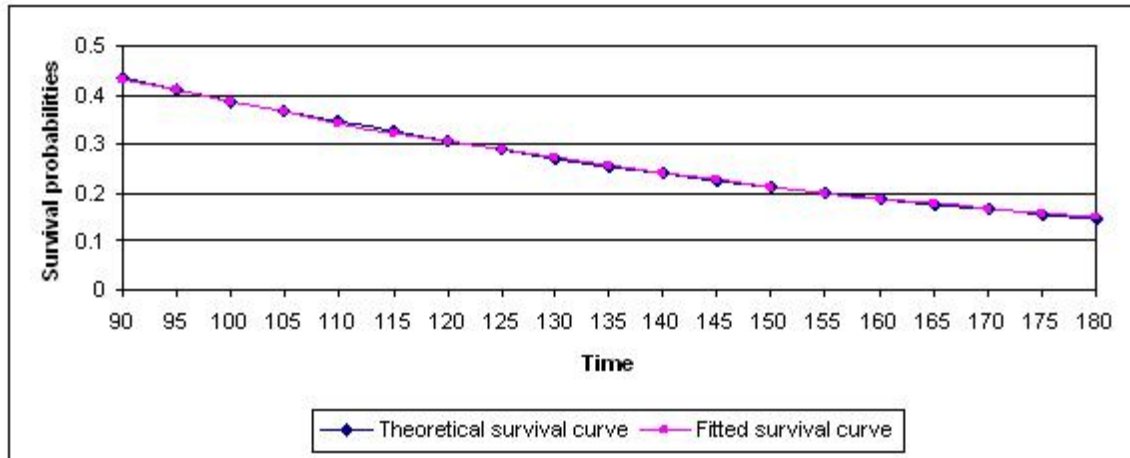


Figure 5.3.8: The theoretical and fitted survival curve for the mixture of exponential and gamma distributions

Mixture of a Weibull distribution and a gamma distribution

The true theoretical survival probability here follows a mixture of a Weibull survival function and a gamma survival function; that is 0.5 of a Weibull survival with a scale parameter equal 0.01 and a shape parameter 1.2 and 0.5 of a gamma survival function with a scale parameter 0.02 and a shape parameter 2. The estimates of the parameters of the extrapolation function are $\hat{\alpha} = 1.000$ and $\hat{\beta} = 0.013$. The theoretical and fitted survival probabilities are presented in table 5.3.9 and displayed in figure 5.3.9.

Table 5.3.9: The theoretical and fitted survival probabilities for the case of the mixture of Weibull and gamma distributions

Time (in months)	The theoretical survival probability	The fitted survival curve probability	The absolute difference
90	0.439	0.440	0.001
95	0.412	0.412	0.000
100	0.387	0.386	0.001
105	0.363	0.362	0.001
110	0.340	0.339	0.001
115	0.319	0.318	0.001
120	0.298	0.298	0.000
125	0.279	0.279	0.000
130	0.261	0.262	0.001
135	0.244	0.245	0.001
140	0.227	0.230	0.003
145	0.212	0.215	0.003
150	0.198	0.202	0.004
155	0.184	0.189	0.005
160	0.172	0.177	0.005
165	0.160	0.166	0.006
170	0.149	0.156	0.007
175	0.139	0.146	0.007
180	0.129	0.137	0.008

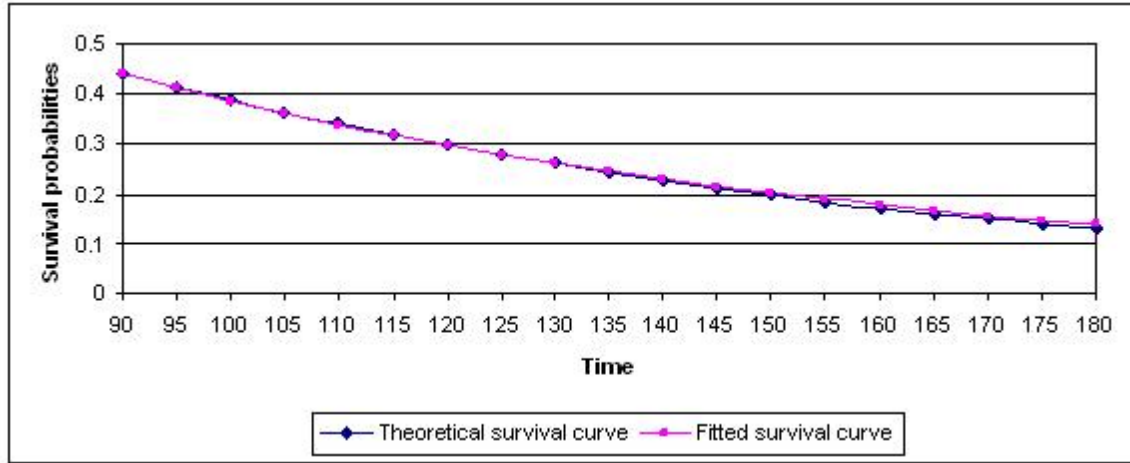


Figure 5.3.9: The theoretical and fitted survival curve for the mixture of Weibull and gamma distributions

5.3.3 Summary of the results of the mathematical check

The proposed extrapolated survival function appears to be suitable for approximating the theoretical survival distributions investigated in the previous section. The theoretical and the fitted survival probabilities are identical to two decimal places and almost identical for three decimal places (table 5.3.1 to table 5.3.9). Looking at the figures that present the theoretical and the fitted survival curves, we see that the proposed extrapolation survival function has given excellent accuracy for the forecasting period which was up to five years (60 months). From table 5.3.10 (on page 95), the maximum norm was less than (or equal to) 0.001 for the 12 months projection period, less than (or equal to) 0.005 for the 36 months projection period and less than (or equal to) 0.007 for the 60 months projection period. The results of the calculation of the maximum norm gave a good indication of the adequacy of the proposed extrapolation function for different scenarios of survival time data distributions. However, the proposed extrapolation survival function becomes less accurate as the forecasting period increases. It might not be wise to forecast for a very long term given the fast change in market dynamics and customer behaviour. The user of the model should decide on the suitable projection period from his/her understanding of the problem at hand.

5.4 The derivation of the standard error of the extrapolation function

Let us consider the estimate of the survival probability obtained from the extrapolation function at t ($t > \tau$) to be

$$y_t = y_t(y_m, \hat{\alpha}, \hat{\beta}) = y_m g(t, \hat{\alpha}, \hat{\beta}) \quad (5.4.1)$$

where $g(t; \hat{\alpha}, \hat{\beta}) = \exp \left\{ \hat{\beta}(\tau - \tau_0)^{\hat{\alpha}} - \hat{\beta}(t - \tau_0)^{\hat{\alpha}} \right\} = g$; $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of α and β that are obtained by solving the minimisation problem

$$\Omega = \sum_{i=k}^m (y_i - y_m g_i)^2 \quad (5.4.2)$$

where $g_i = g_i(t_i; \hat{\alpha}, \hat{\beta}) = \exp \left\{ \hat{\beta}(\tau - \tau_0)^{\hat{\alpha}} - \hat{\beta}(t_i - \tau_0)^{\hat{\alpha}} \right\}$, $y_i = S(t_i)$ are Kaplan-Meier estimates ($t_i \leq \tau$), $\hat{\alpha}$ and $\hat{\beta}$ are functions of the random variables $y_i : i = k, \dots, m$. That is, the estimate y_t of the extrapolation function is a function of $y_m, \hat{\alpha}, \hat{\beta}$.

Our objective now is to find the standard error of y_t using the variance-covariance matrix of $y_m, \hat{\alpha}, \hat{\beta}$. Let us say $y_t = f(y_m, \hat{\alpha}, \hat{\beta})$, then the variance of y_t is equal to $\text{var}(f(y_m, \hat{\alpha}, \hat{\beta}))$. We use the delta method to find a linear approximation for $f(y_m, \hat{\alpha}, \hat{\beta})$ and thereafter the variance of this linear approximation (Oehlert, 1992). Using the Taylor expansion of first order

$$\begin{aligned} f(y_m, \hat{\alpha}, \hat{\beta}) &\approx f(\eta_m, \alpha, \beta) + (y_m - \eta_m) \frac{\partial f(y_m, \hat{\alpha}, \hat{\beta})}{\partial y_m} \\ &\quad + (\hat{\alpha} - \alpha) \frac{\partial f(y_m, \hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} + (\hat{\beta} - \beta) \frac{\partial f(y_m, \hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} \end{aligned}$$

That is

$$\begin{aligned}
 f(y_m, \hat{\alpha}, \hat{\beta}) &\approx C + \begin{pmatrix} y_m - \eta_m \\ \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix}^T \begin{pmatrix} \frac{\partial f}{\partial y_m} \\ \frac{\partial f}{\partial \alpha} \\ \frac{\partial f}{\partial \beta} \end{pmatrix} \\
 &\approx C + (\hat{\theta} - \theta)^T \left(\frac{\partial f}{\partial \theta} \right)
 \end{aligned} \tag{5.4.3}$$

$\hat{\theta} = [y_m, \hat{\alpha}, \hat{\beta}]^T$, $\theta = [\eta_m, \alpha, \beta]^T$, C constant.

$$\begin{aligned}
 \text{var}(y_t) &= \text{var} \left[f(y_m, \hat{\alpha}, \hat{\beta}) \right] = \text{var} \left(C + [\hat{\theta} - \theta]^T \left[\frac{\partial f}{\partial \theta} \right] \right) \\
 &= \left[\frac{\partial f}{\partial \theta} \right]^T \text{var}(\hat{\theta} - \theta) \left[\frac{\partial f}{\partial \theta} \right] \\
 &= \left[\frac{\partial f}{\partial \theta} \right]^T \text{var}(\hat{\theta}) \left[\frac{\partial f}{\partial \theta} \right]
 \end{aligned} \tag{5.4.4}$$

$$\text{var}(\hat{\theta}) = \text{var} \begin{pmatrix} y_m \\ \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \text{var}(y_m) & \text{cov}(y_m, \hat{\alpha}) & \text{cov}(y_m, \hat{\beta}) \\ \text{cov}(\hat{\alpha}, y_m) & \text{var}(\hat{\alpha}) & \text{cov}(\hat{\alpha}, \hat{\beta}) \\ \text{cov}(\hat{\beta}, y_m) & \text{cov}(\hat{\beta}, \hat{\alpha}) & \text{var}(\hat{\beta}) \end{pmatrix} \tag{5.4.5}$$

That is, to calculate the standard error of y_t one has to substitute 5.4.5 in 5.4.4 and take the square root of the right hand side of equation 5.4.4. The $\text{var}(y_m)$ in equation 5.4.5 can be obtained from the Greenwood formula, but to find $\text{cov}(y_m, \hat{\alpha})$, $\text{cov}(y_m, \hat{\beta})$, $\text{cov}(\hat{\alpha}, \hat{\beta})$, $\text{var}(\hat{\alpha})$ and $\text{var}(\hat{\beta})$, one has to write $\hat{\alpha}$ and $\hat{\beta}$ as a linear function of $y_{i,s}$. To do so, we take the linear approximation of Ω_α and Ω_β , where $\Omega_\alpha = \frac{\partial \Omega}{\partial \alpha}$, $\Omega_\beta = \frac{\partial \Omega}{\partial \beta}$. That is:

$$\Omega_\alpha \approx \Omega_\alpha(\alpha, \beta) + (\hat{\alpha} - \alpha)\Omega_{\alpha\alpha}(\hat{\alpha}, \hat{\beta}) + (\hat{\beta} - \beta)\Omega_{\alpha\beta}(\hat{\alpha}, \hat{\beta})$$

$$\Omega_\beta \approx \Omega_\beta(\alpha, \beta) + (\hat{\alpha} - \alpha)\Omega_{\beta\alpha}(\hat{\alpha}, \hat{\beta}) + (\hat{\beta} - \beta)\Omega_{\beta\beta}(\hat{\alpha}, \hat{\beta})$$

It follows that:

$$\begin{pmatrix} \Omega_\alpha(\hat{\alpha}, \hat{\beta}) \\ \Omega_\beta(\hat{\alpha}, \hat{\beta}) \end{pmatrix} \approx \begin{pmatrix} \Omega_\alpha(\alpha, \beta) \\ \Omega_\beta(\alpha, \beta) \end{pmatrix} + D \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} \quad (5.4.6)$$

where

$$D = \begin{pmatrix} \Omega_{\alpha\alpha}(\alpha, \beta) & \Omega_{\alpha\beta}(\alpha, \beta) \\ \Omega_{\beta\alpha}(\alpha, \beta) & \Omega_{\beta\beta}(\alpha, \beta) \end{pmatrix}$$

$$\Omega_{\alpha\alpha} = \frac{\partial \Omega_\alpha}{\partial \alpha}, \Omega_{\alpha\beta} = \frac{\partial \Omega_\alpha}{\partial \beta}, \Omega_{\beta\alpha} = \frac{\partial \Omega_\beta}{\partial \alpha}, \Omega_{\beta\beta} = \frac{\partial \Omega_\beta}{\partial \beta}$$

The second derivatives of Ω given in D are evaluated at their observed expectations, so they will be treated as constant. From equation 5.4.6

$$\text{var} \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} = \text{var} \left[D^{-1} \begin{pmatrix} \Omega_\alpha(\hat{\alpha}, \hat{\beta}) \\ \Omega_\beta(\hat{\alpha}, \hat{\beta}) \end{pmatrix} \right]$$

Hence

$$\begin{aligned} \text{var} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} &= D^{-1} \text{var} \left[\begin{pmatrix} \Omega_\alpha(\hat{\alpha}, \hat{\beta}) \\ \Omega_\beta(\hat{\alpha}, \hat{\beta}) \end{pmatrix} \right] [D^{-1}]^T \\ &= D^{-1} \begin{pmatrix} \text{var} [\Omega_\alpha(\hat{\alpha}, \hat{\beta})] & \text{cov} [\Omega_\alpha(\hat{\alpha}, \hat{\beta}), \Omega_\beta(\hat{\alpha}, \hat{\beta})] \\ \text{cov} [\Omega_\beta(\hat{\alpha}, \hat{\alpha}), \Omega_\beta(\hat{\alpha}, \hat{\beta})] & \text{var} [\Omega_\beta(\hat{\alpha}, \hat{\beta})] \end{pmatrix} [D^{-1}]^T \end{aligned} \quad (5.4.7)$$

From 5.4.7 we learned that the problem now is how to calculate the variances and covariances of Ω_α and Ω_β

$$\Omega_\alpha = \frac{\partial \Omega}{\partial \alpha} = \frac{\partial \sum_{i=k}^m (y_i - y_m g_i)^2}{\partial \alpha} = -2 \sum_{i=k}^m [y_i - y_m g_i] [y_m g_{i\alpha}] \quad (5.4.8)$$

where $g_{i\alpha} = \frac{\partial g_i}{\partial \alpha}$. To linearise Ω_α we linearise each term in the summation, i.e., we linearise

$$h(y_i, y_m) = [y_i - y_m g_i] y_m g_{i\alpha}.$$

The linear approximation of $h(y_i, y_m)$ is expressed as:

$$\begin{aligned} h(y_i, y_m) &\approx h(\eta_i, \eta_m) + (y_i - \eta_i) \frac{\partial h}{\partial y_i} + (y_m - \eta_m) \frac{\partial h}{\partial y_m} \\ &= C_1 + (y_i - \eta_i)(\eta_m g_{i\alpha}) + (y_m - \eta_m)[\eta_i g_{i\alpha} - 2\eta_m g_i g_{i\alpha}] \\ &= C_2 + y_i \eta_m g_{i\alpha} + [\eta_i g_{i\alpha} - \eta_m g_i g_{i\alpha}] y_m \end{aligned} \quad (5.4.9)$$

where C_1 and C_2 are constants, and $\eta_i = E(y_i)$. Then

$$\Omega_\alpha \approx \text{constant} - 2 \left[\sum_{i=k}^{m-1} y_i \eta_m g_{i\alpha} + \sum_{i=k}^{m-1} (\eta_i g_{i\alpha} - 2\eta_m g_i g_{i\alpha}) y_m \right] \quad (5.4.10)$$

Thereafter, Ω_α can be written as

$$\Omega_\alpha \approx a_1 y_1 + a_2 y_2 + \dots + a_m y_m \quad (5.4.11)$$

where a_i is the coefficient of y_i in Ω_α

In the same manner Ω_β can be written as follows:

$$\Omega_\beta \approx \text{constant} - 2 \left[\sum_{i=k}^{m-1} y_i \eta_m g_{i\beta} + \sum_{i=k}^{m-1} (\eta_i g_{i\beta} - 2\eta_m g_i g_{i\beta}) y_m \right] \quad (5.4.12)$$

In the same way, we can linearise Ω_β and write it as follows:

$$\Omega_\beta \approx b_1 y_1 + b_2 y_2 + \dots + b_m y_m \quad (5.4.13)$$

where b_i is the coefficient of y_i in the linearised version of Ω_β .

In matrix format

$$\begin{pmatrix} \Omega_\alpha \\ \Omega_\beta \end{pmatrix} = CY \quad (5.4.14)$$

where $C = \begin{pmatrix} a_1 & a_2 & \dots & a_m \\ b_1 & b_2 & \dots & b_m \end{pmatrix}$ and $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$

Using 5.4.7 and 5.4.14

$$\text{var} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = D^{-1} C \text{var}(Y) C^T (D^{-1})^T \quad (5.4.15)$$

If the elements of D^{-1} are d_{ij} $i, j = 1, 2$ then $\hat{\theta}$ can be written as follows

$$\hat{\theta} = \begin{pmatrix} y_m \\ \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & d_{11} & d_{12} \\ 0 & d_{21} & d_{22} \end{pmatrix} \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ a_1 & a_2 & \dots & a_{m-1} & a_m \\ b_1 & b_2 & \dots & b_{m-1} & b_m \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix} \quad (5.4.16)$$

The covariance matrix in 5.4.5 can be directly calculated from 5.4.16 given that

$$\text{cov}(y_i, y_j) = y_i y_j \sum_{i \leq j} \frac{d_i}{n_i(n_i - d_i)} \quad (5.4.17)$$

where $i \leq j$. The Greenwood's formula is a special case of 5.4.17 (Jewell *et al.*, 2005). It is important to mention that this method of approximation works well if we have a large sample size. With a small sample, the linearization might not be a good idea.

In the next section, we apply the proposed extrapolation function and the derived standard error on the data set defined in chapter four to obtain a 95% confidence interval for survival probabilities beyond the last observed failure time τ . Code for the proposed extrapolation function (with its standard error) is written in Octave and available to both the scientific community and business practitioners (see Appendix A).

5.5 The conditional survival function

Let us assume that at time t_0 we have in the system N clients who have survived for times t_1, t_2, \dots, t_N . The task is to calculate the conditional survival time that a customer will survive over the interval $t_0 \rightarrow t_0 + \Delta$; where Δ is specified in months ($\Delta > 0$), for example $\Delta = 12$ months. The conditional survival probability for client j is then defined as follows:

$$\hat{S}_j(t_j + t/t_j) = \begin{cases} \frac{\hat{S}(t_j+t)}{\hat{S}(t_j)}, & 0 < t \leq \Delta \\ 0, & t > \Delta \end{cases} \quad (5.5.1)$$

where

$$\hat{S}_j(t_j + t) = \begin{cases} \hat{S}_{KM}, & 0 < t_j + t \leq \tau \\ \hat{S}_{Ext}, & t_j + t > \tau \end{cases} \quad (5.5.2)$$

\hat{S}_{KM} is the Kaplan-Meier estimate of survival probability of customer j at time $t_j + t$.

\hat{S}_{Ext} is the extrapolation function estimate of survival probability of customer j at time $t_j + t$.

5.6 Applications

5.6.1 Estimation of future survival probabilities

This section shows how the proposed survival function works. For this purpose, the variable age group was used as an example. In each age group, the survival probability was obtained from Kaplan-Meier and, thereafter, the Kaplan-Meier survival probabilities were used to estimate the parameters of the extrapolation function. Customer survival time probabilities, for a 12-month forecasting period beyond the last observed failure time, were calculated and the standard error of the estimate is attached.

Figure 5.6.1 shows a 95% confidence band of the survival probabilities for customers of age less than 26 years. The estimates before the vertical dashed line is obtained

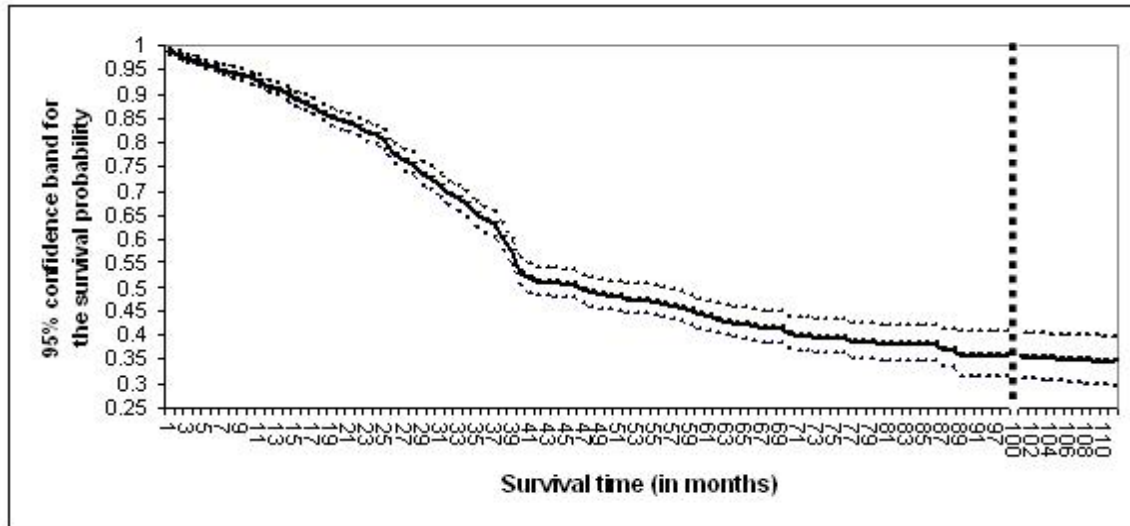


Figure 5.6.1: Confidence band of the survival probabilities for customers of age less than 26 years

from Kaplan-Meier and estimates beyond the vertical dashed line are obtained from the proposed extrapolation survival function.

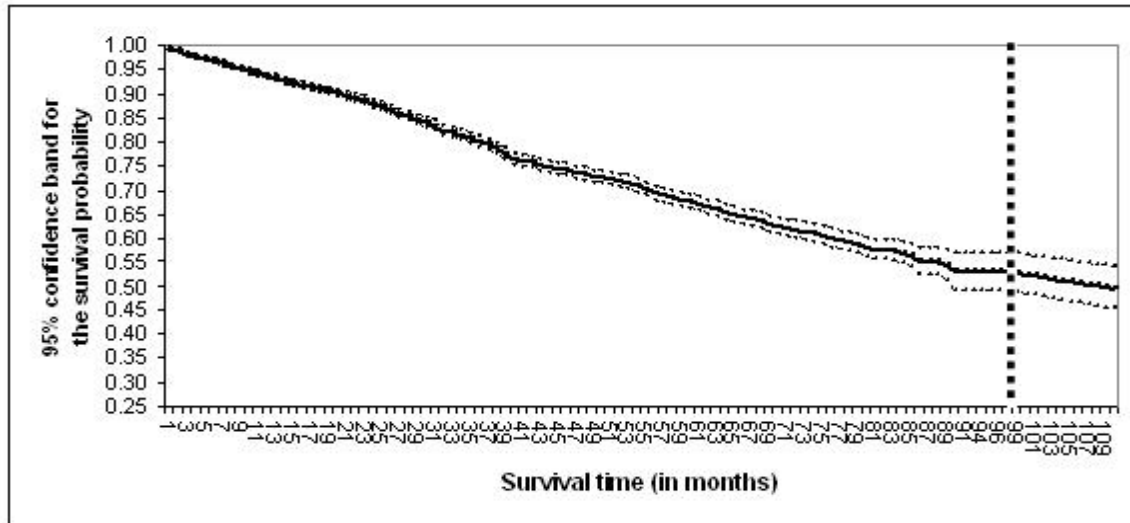


Figure 5.6.2: Confidence band of the survival probabilities for customers of age between 26 and 40 years

Figure 5.6.2 shows a 95% confidence band of the survival probabilities for customers between 26 and 40 years of age. The estimates before the vertical dashed line are obtained from Kaplan-Meier and estimates beyond the vertical dashed line are obtained from the proposed extrapolation survival function.

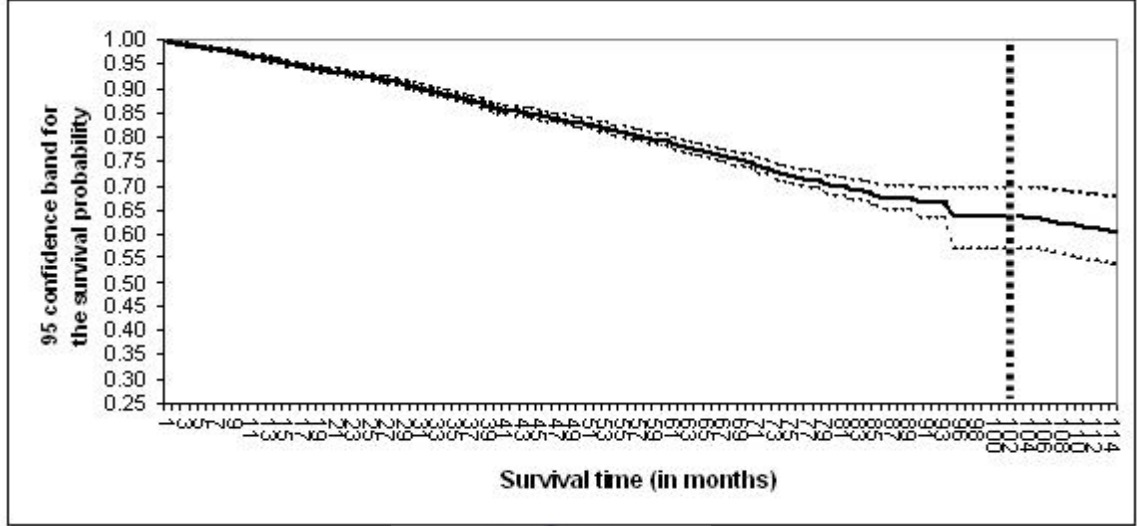


Figure 5.6.3: Confidence band of the survival probabilities for customers of age more than 40 years

Figure 5.6.3 shows a 95% confidence band of the survival probabilities for customers more than 40 years of age. The estimates before the vertical dashed line are obtained from Kaplan-Meier and estimates beyond the vertical dashed line are obtained from the proposed extrapolation survival function.

5.6.2 Establishing confidence limits for customer lifetime value

In this section we use survival probabilities that were calculated from the extrapolation function to obtain the customer lifetime value (CLV). Moreover, we derive the standard error of the CLV using the delta method (Oehlert, 1992). To illustrate this application we consider the lifetime value model that has been proposed by Lu (2003). This model is presented in chapter two in equation 2.2.3. We redefine this model to enable us to project the lifetime value for the period τ to $\tau + \Delta$. Once again, τ is the last observed failure time and Δ is given in months. This can be expressed as follows:

$$CLV = m \sum_{t=\tau+1}^{\tau+\Delta} \frac{y_t}{(1 + d/12)^t} \quad (5.6.1)$$

where m is the customer monthly margin, y_t is an estimate of customer survival probabilities S_t and is obtained from the extrapolation function, and d is the discount rate. We rewrite the above equation as follows:

$$CLV = m \sum_{t=\tau+1}^{\tau+\Delta} a_t y_t = m A^T Y_t \quad (5.6.2)$$

where $a_t = (1 + d/12)^t$, A is a vector of a_t , Y_t is a vector of y_t : $t = \tau + 1$ to $\tau + \Delta$. It is obvious that for a given m and d , the calculation of the standard error of CLV involve only the calculation of the variance-covariance matrix of Y_t . That is:

$$\text{var}(CLV) = \text{var}(m A^T Y_t) = m^2 A^T \text{var}(Y_t) A \quad (5.6.3)$$

Using the delta method (Oehlert, 1992) we can calculate the variance-covariance matrix of Y_t . This can be done by linearizing each element y_t of the vector Y_t . Each of the elements of Y_t is a function of $y_m, \hat{\alpha}, \hat{\beta}$. We denote y_t by $f(t, y_m, \hat{\alpha}, \hat{\beta})$. Then

$$\begin{aligned} f(t, y_m, \hat{\alpha}, \hat{\beta}) &\approx f(t, \eta_m, \alpha, \beta) + (y_m - \eta_m) \frac{\partial f(t, y_m, \hat{\alpha}, \hat{\beta})}{\partial y_m} \\ &+ (\hat{\alpha} - \alpha) \frac{\partial f(t, y_m, \hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} + (\hat{\beta} - \beta) \frac{\partial f(t, y_m, \hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} \end{aligned}$$

for $t = \tau + 1$ to $\tau + \Delta$. In this way we obtain Δ equations of which each is of the above format. Following the same argument as in equation 5.4.3 and 5.4.4 the variance-covariance matrix of Y_t can be expressed as follows:

$$\text{var}(Y_t) = F \text{var}(\hat{\theta}) F^T \quad (5.6.4)$$

where F is a Δ by 3 matrix of which each row represents the derivatives of y_t with respect to $\hat{\theta}$, $\hat{\theta} = [y_m, \hat{\alpha}, \hat{\beta}]^T$, $\theta = [\eta_m, \alpha, \beta]^T$. The variance-covariance matrix of $\hat{\theta}$ can be calculated exactly in the same way as we did in section 5.4. Thereafter, we

substitute 5.6.4 in 5.6.3 to get variance of CLV. For a large sample size, 95% confidence limits for CLV will be:

$$mA^TY_t \pm 1.96 * m\sqrt{A^TF\text{var}(\hat{\theta})F^TA} \quad (5.6.5)$$

An Octave code is provided in the appendix to perform the computation of the above equations.

Example:

Suppose that we have a customer of age more than 40 years who survived exactly τ months at time τ . Let us also assume a discount factor of 10%, a constant monthly margin of \$100 and $\Delta = 12$ months. We would like to calculate the lifetime value of this customer using the above equations. This will result is a customer lifetime value equal to \$707.02 with standard error \$39.04. Using the normal approximation, we obtain a 95% confidence interval of (630.51, 780.53).

Customer equity of a firm can be directly obtained from customer lifetime value because it is the aggregate of the value of all customers. Note that if at the time of estimation there was a customer who survived less than τ months then the survival probabilities of this customer has to be obtained from the Kaplan-Meier estimate for $t \leq \tau$ and the corresponding variance-covariance matrix can be obtained using equation 5.4.17.

5.7 Summary and conclusion

This chapter dealt with one of the important issues in the customer survival time estimation problem; that is, the estimation of the survival probabilities and the survival time beyond the empirical data. The practical motivation for extrapolating the survival curve beyond the empirical distribution originates from two issues, that of calculating survival probabilities (retention rate) beyond the empirical data and of calculating expected survival time. In this regard we proposed a function that can be

used to extrapolate the survival probabilities beyond the last observed failure time. The estimation of parameters of the extrapolation function is based completely on the Kaplan-Meier estimate of the survival probabilities. The mathematical accuracy of the proposed function was checked against various theoretical lifetime data distributions and a mixture of life time data distributions. The theoretical and the fitted survival probabilities are identical to two decimal places and almost identical for three decimal places. The proposed extrapolation survival function has given excellent accuracy for the forecasting period which was up to five years (60 months). The maximum norm was less than (or equal to) 0.001 for a 12-month projection period, less than (or equal to) 0.005 for a 36-month projection period and less than (or equal to) 0.007 for a 60-month projection period. The results of the calculation of the maximum norm gave a good indication of the adequacy of the proposed extrapolation function for different scenarios of survival time data distributions. However, the proposed extrapolation survival function becomes less accurate as the forecasting period increases. It might not be wise to forecast for a very long term given the rapid changes in market dynamics and customer behaviour. The user of the model should decide on the suitable projection period from his/her understanding of the problem at hand.

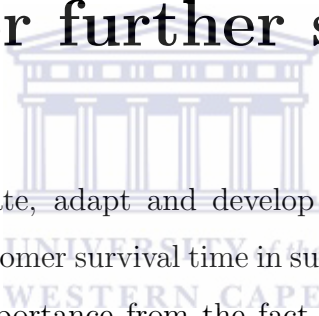
After having the mathematical error checked, we derived the standard error of the estimate of the extrapolation function. The derivation is based on the delta method. Finally we implemented the proposed model on a real data set to estimate a 95% confidence band for the extrapolated survival probabilities for a projection period of 12 months after the last observed failure time. Numerical and visual examinations suggest that the proposed survival function works very well. The code is written in Octave and is ready for use by business managers where the objective is to enhance customer retention and to emphasise a customer-centric approach. Although we have developed this model to serve the business community, it can be applied and used beyond the customer survival time data to cover clinical trial applications.

Table 5.3.10: The maximum norms for different projection periods and different theoretical distributions

The scenario of the theoretical distribution	The maximum norm for different projection period		
	12 months	36 months	60 months
Gamma distribution with scale parameter 0.02 and shape parameter 2	0.000	0.001	0.002
Mixture of two exponentials (1): 0.9 of exponential with scale parameter 0. plus 0.01 of exponential with scale parameter 0.02	0.000	0.002	0.002
Mixture of two exponentials (2): 0.5 of exponential with scale parameter 0.01 plus 0.5 of exponential with scale parameter 0.02	0.000	0.000	0.000
Mixture of two Weibulls (1): 0.9 of Weibull with scale parameter 0.005 and shape parameter 1.2 plus 0.1 Weibull with scale parameter 0.005	0.001	0.002	0.003
Mixture of two Weibulls (2): 0.5 of Weibull with scale parameter 0.01 and shape parameter 1.2 plus 0.5 Weibull with scale parameter 0.008 and shape parameter 1.5	0.001	0.004	0.007
Mixture of two gammas: 0.5 of gamma with scale parameter 0.02 and shape parameter 2 plus 0.5 of gamma with scale parameter 0.03 and shape parameter 3	0.001	0.005	0.007
Mixture of exponential and Weibull: 0.5 of exponential with scale parameter 0.01 plus 0.5 of Weibull with scale parameter 0.01 and shape parameter 1.5	0.000	0.000	0.000
Mixture of exponential and gamma: 0.5 of exponential with scale parameter 0.01 plus 0.5 of gamma with scale parameter 0.02 and shape parameter 2	0.001	0.002	0.003
Mixture of Weibull and gamma: 0.5 of Weibull with scale parameter 0.01 and shape parameter 1.2 plus 0.5 of gamma scale parameter 0.02 and shape parameter 2	0.003	0.005	0.007

Chapter 6

Conclusion, recommendations and directions for further studies



This study aimed to illustrate, adapt and develop methods of survival analysis in analysing and estimating customer survival time in subscription-based businesses. This area of research gains its importance from the fact that customer survival time is a crucial element in the process of implementing a customer-centric approach, mainly in estimating a firm's value based on their customers value. This involves managing acquisition and retention, justifying and testing the return on investment in a customer-centric approach and, generally, maintaining and making customer-firm relationships more valuable and profitable. Two main objectives were set: The first objective was to redefine the existing survival analysis techniques - that were mainly used to solve questions related to medical field - in business terms and to discuss their uses in order to understand various issues related to the customer-firm relationships, while the second objective was to extrapolate the customer survival curve beyond the empirical distribution.

In relation to the redefinition of the current survival analysis techniques to meet the business needs, we presented the basic formulation of survival and hazard functions in business terms. The underlying assumption about the distribution of customer survival time and the differences between survival analysis in the medical field and business

field due to the nature of business data and business needs are discussed. We have also discussed and shown the applicability of survival analysis techniques - including non-parametric, semi-parametric and parametric models - in understanding and analysing various issues related to the customer-firm relationship. A particular data set from a well-established subscription-based business company was used to investigate the ability of current survival analysis techniques in understanding and managing various issues related to the customer-firm relationship. The data set contains demographic and usage-related variables. We highlighted the business and the methodological insights in the process of applying survival analysis techniques. This includes identification of the significant variables that affect customer relationships with a firm using a stratified version of the Cox regression model, modelling and understanding customer loyalty and churn using the hazard ratio obtained from the Cox regression model, the cumulative hazard obtained from Nelson-Aalen, and survival probabilities calculated from Kaplan-Meier; the use of the Kaplan-Meier method to compare different marketing campaigns and different customer groups; and the segment-based estimation of customer mean survival time using the non-parametric method of Kaplan-Meier and the parametric method using exponential, Weibull and gamma distributions. Generally, a direct application of current survival analysis techniques to analyse customer survival time and customer risk of cancellation has shown great potential.

In another direction, the application of current techniques in estimating customer mean survival time was studied. The results on the estimate for customer mean survival time are obtained via different nonparametric and parametric approaches; namely, the Kaplan-Meier method as well as the exponential, Weibull and gamma regression models. The estimate of the mean from Kaplan-Meier method, exponential regression, Weibull regression, and gamma regression varies greatly. This suggests that a careful investigation of the method used to extend the survival curve and the choice of the parametric model are extremely important, especially in the type of business problems where a high degree of censoring is expected. For example, customers with ages less than 26 years have the lowest variability in the estimates of mean survival time when

different methods of estimation were used; it has to be noted that this age group has the lowest degree of censoring.

We suggest that in estimating the customer mean survival time, one would prefer non-parametric methods with a careful plan to deal with extrapolation issues. However, from our understanding of the fast dynamics of the market and the fast change in customer behaviour, a conditional mean survival time based on the empirical distribution of the data - for each sub-population and for a reasonable time horizon - will be the way forward. This will enable us to get accurate and useful inferences. Therefore, we dedicate a crucial part of this research to the issue of extrapolating the survival curve and we believe that this is where this study is making a major contribution to the methodology and the literature of survival analysis.

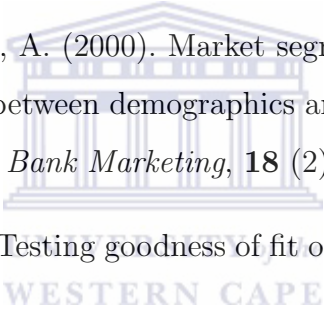
In relation to the extrapolation of the survival curve we propose a survival function that can be used to extrapolate the survival probabilities beyond the empirical data for projection purposes. The estimation of the parameters of the extrapolation function is based completely on the Kaplan-Meier estimate of the survival probabilities in the observation period. The mathematical accuracy of the proposed function was checked against various theoretical lifetime data distribution and a mixture of lifetime data distributions. The theoretical and the fitted survival probabilities are identical to two decimal places and almost identical for three decimal places. The proposed extrapolation survival function has given excellent accuracy for a forecasting period of five years (60 months). The maximum norm was less than (or equal to) 0.001 for a 12-month projection period, less than (or equal to) 0.005 for a 36-month projection period and less than (or equal to) 0.007 for a 60-month projection period. However, the proposed extrapolation survival function became less accurate as the forecasting period increased; this suggests that it might not be wise to forecast for a very long term given the fast change in market dynamics and customer behaviour. The user of the model should decide on a suitable projection period from his/her understanding of the problem faced. After having the mathematical error checked, we derived the standard error of the estimates of the extrapolation function using the delta method. Finally,

we implemented the proposed model on a real data set to estimate a 95% confidence band for the extrapolated survival probabilities for a projection period of 12 months after the last observed failure time. Numerical and visual examination suggest that the proposed survival function works very well. The code is written in Octave and is ready to be used of business managers where the objective is to enhance customer retention and emphasise a customer-centric approach. Although we have developed this model to serve the business community, it can be applied and used beyond the customer survival time data to cover clinical trial applications.

In the last part of the study, we developed an expression for conditional survival time probabilities using the Kaplan-Meier survival probability together with the survival probability obtained from the extrapolation survivor function.

It is important that the extrapolation survival function accounts for future changes in customer behaviour. The estimation of distant probabilities, based on historical data, might not be accurate enough to produce accurate forecasting. This is due to the fast change in customer behaviour and market dynamics (Zeihaml *et al.*, 2006). For further studies, we suggest that researchers should investigate extrapolation models that are based on both Kaplan-Meier estimates and a function to represent the change in customer behaviour over time.

Bibliography

- 
- [1] Aalen, O. O. (1978). Nonparametric inference family counting processes. *Annals of Statistics*, **6** (2): 701-726.
- [2] Alfansi, L. and Sargeant, A. (2000). Market segmentation in Indonesian banking sector: the relationship between demographics and the desired customer benefits. *International Journal of Bank Marketing*, **18** (2): 64-74.
- [3] Anderson, P. K. (1982). Testing goodness of fit of Cox's regression and life model, *Biometrics*. **38**: 67-77.
- [4] Andronikidis, A. I. and Dimitriadis, N. I. (2003). Segmentation of bank customers by utilizing ethnic/cultural profile dimension: a quantitative approach. *Journal of Marketing Management*, **19**: 629-655.
- [5] Badgett, M. and Stone, M. (2005). Multidimensional segmentation at work: Driving an operational model that integrates customer segmentation with customer management. *Journal of Targeting, Measurement and Analysis for Marketing*, **13** (2): 103-121.
- [6] Bell, D., Deighton, J., Reinartz, W. J., Rust, R. T. and Swartz, G. (2002). Seven barriers to customer equity management. *Journal of Services Research*, **5** (1): 77-85.
- [7] Berger, P. D. and Nasr, N. I. (1998). Customer lifetime value: marketing models and applications. *Journal of interactive marketing*, **12**(1): 17-30.

- [8] Berger, P. D., Bolton, R. N., Bowman, D., Briggs, E., Kumar, V., Parasuraman, A. and Terry, C. (2002). Marketing actions and the value of customer assets: a framework for customer asset management. *Journal of Services Research*, **5** (1): 39-54.
- [9] Blattberg, R. and Deighton, J. (1996). Manage marketing by the customer equity test. *Harvard Business Review*, **74** (4): 136-144.
- [10] Boggs, P.T., Tolle, J. W. and Wang, P. (1982). On local convergence of quasi-newton methods for constrained optimization problem. *SIAM Journal on Control and Optimization*, **20** (2): 161-171.
- [11] Bolton, R. N., Lemon, K. N. and Verhoef, P. C. (2004). The theoretical underpinning of customer asset management: a framework and propositions for future research. *Journal of the Academy of Marketing Science*, **32** (3): 271-292.
- [12] Bonnons, J. F., Painer, E. R., Titts, A. L. and Zhou, J. L. (1992). Avoiding the Maratos effect by means of monotone line search. II. Inequality constrained problems - feasible iterates, *SIAM Journal on Numerical Analysis*, **29** (4): 1187-1202.
- [13] Bruwer, J. and Elton, L. (2007). Wine-related lifestyle (WRL) market segmentation: demographic and behavioral factors. *Journal of Wine Research*, **18** (1): 19-34.
- [14] Campbell, D. and Frei, F. (2004). The persistence of customer profitability: Empirical evidence and implications from a financial services firm. *Journal of Services Research*, **7**(2): 107-123.
- [15] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, **34** (2): 187-220.
- [16] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62** (2): 269-276.

- [17] Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*. Florida: Chapman & Hall/CRC.
- [18] Driver, C. and Johnston, R. (2001). Understanding service customers: the value of hard and soft attributes. *Journal of Services Research*, **4**: 130-139.
- [19] Encinas, N., Alfonso, D., Alvarez, C., Perez-Navarro, A. and Garcia-Franco, F. (2007). Energy market segmentation for distributed energy resources implementation purposes. *IET Generation, Transmission & Distribution*, **1** (2): 324-330.
- [20] Fader, P. S., Hardie, B. G. S. and Lee, K. L. (2005). RFM and CLV: Using ISO-CLV curves for customer base analysis. *Journal of Marketing Research*, **42** (4): 415-430.
- [21] Figini, S., Giudici, P. and Brooks, S. P. (2007). Bayesian feature selection to estimate customer survival. Working paper. [Available: <http://www.statslab.cam.ac.uk/~steve/mypapers/figgb06.pdf>] (read in May 2007).
- [22] Figini, S. and Giudici, P. (2007a). Statistical models for customer lifetime value. Working paper. [Available: <http://www.unipv.it/dipstea/wp/32.pdf>] (read in May 2007).
- [23] Figini, S. and Giudici, P. (2007b). Bayesian models to estimate customer survival. Working paper. [Available: <http://www.unipv.it/dipstea/wp/29.pdf>] (read in May 2007).
- [24] Figini, S. (2007). Bayesian penalized models to estimate customer survival, working paper, [Available: <http://www.unipv.it/dipstea/wp/34.pdf>] (read in May 2007).
- [25] Gopalan, R. (2007). Customer portfolio management using Z-ranking of customer segments and the LTV perturbation method. *Database Marketing and Customer Strategy Management*, **14** (3): 225-235.

- [26] Gouthier, M. and Schmid, S. (2003). Customers and customer relationships in the service firms: the perspective of the resource-based view. *Journal of Marketing Theory*, **3** (1): 119-143.
- [27] Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests in diagnostics based on weighted residuals. *Biometrika*. **81** (3): 515-526.
- [28] Gross, A. J. and Clark, V. A. (1975). *Survival distributions: reliability applications in the biomedical sciences*. New York: John Wiley and Sons.
- [29] Gupta, S. and Lehmann, D. R. (2003). Customer as assets. *Journal of Interactive Marketing*, **17** (1): 9-24.
- [30] Gupta, S., Lehmann, D. R. and Stuart, J. A. (2004). Valuing customers, *Journal of marketing Research*, **41** (1): 7-18.
- [31] Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N. and Sriram, S. (2006). Modeling customer lifetime value, *Journal of Services Research*, **9** (2): 139-155.
- [32] Han, S. P. (1976). Superlinearly convergent variable metric algorithm for general nonlinear programming problems. *Mathematical Programming*, **11**: 263-282.
- [33] Hogan, J. E., Lehmann, D. R., Merino, M., Srivastava, R. K., Thomas, J. S. and Verhoef, P. C. (2002). Linking customer assets to financial performance. *Journal of Services Research*, **5** (1): 26-38.
- [34] Hogan, J. E., Lemon, K. N. and Libai, B. (2003). What is the true value of a lost customer? *Journal of Services Research*, **5** (3): 196-208.
- [35] Hosmer, D. W., Jr. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression modeling of time to event data*. John Wiley and Sons: New York.
- [36] Hougaard, P. (2001). *Analysis of multivariate survival data*. Springer: New York.

- [37] Jain, D. and Singh, S. S. (2002). Customer lifetime value research in marketing: a review and future directions. *Journal of Interactive Marketing*, **16** (2): 34-46.
- [38] Jenkins, S. P. (2005). Survival analysis. [Available: <http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/pdfs/ec968lnotesv6.pdf>] (read in August 2005).
- [39] Jewell, N. P., Lei, X., Ghani, A. C., Donnelly, C. A., Leung, G. M., Ho, L. M., Cowling, B. and Hedley, A. J. (2005). Nonparametric estimation of the case fatality ratio with competing risks data: an application to severe acute respiratory syndrome (SARS). Berkeley Electronic Press: University of California, Berkeley. [Available: <http://www.bepress.com/ucbbiostat/paper176/>] (read in October 2007).
- [40] Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. **53** (282): 457-481.
- [41] Kim, S., Jung, T., Suh, E. and Hwang, H. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert systems with applications*, **31** (1): 106-107.
- [42] King, J. T. Jr, Justice, A. C., Roberts, M. S., Chang, C. C. and Fusco, J.S. (2003). Long-term HIV/AIDS survival estimation in the highly active antiretroviral therapy era. *Medical Decision Making*. **23** (1): 9-20.
- [43] Klein, J. P and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.
- [44] Kumar, V., Ramani, G. and Bohling, T. (2004). Customer lifetime value: approaches and best practice applications. *Journal of Interactive Marketing*, **18** (3): 60-72.

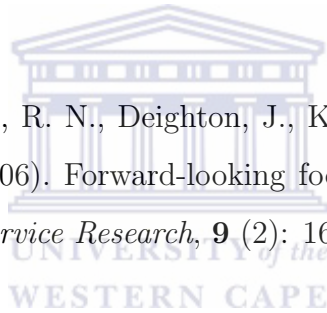
- [45] Kumar, V. and Petersen, J. A. (2005). Using customer-level marketing strategy to enhance firm performance: a review of theoretical and empirical evidence. *Journal of the Academy of Marketing Science*, **33** (4): 504-519.
- [46] Kumar, V., Lemon, K. N. and Parasuraman, A. (2006). Managing customers for value: An overview and research agenda. *Journal of Services Research*, **9** (2): 87-94.
- [47] Kumar, V. and Reinartz, W. (2006). Knowing what to sell, when, and to whom. *Harvard Business Review*, **84** (3): 131-137.
- [48] Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, **35**: 139-156.
- [49] Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. New Jersey: John Wiley and Sons.
- [50] Lee, E. T. and Wang, J. W. (2003). *Statistical methods for survival data analysis*. New Jersey: John Wiley and Sons.
- [51] Lev, B. and Zarowin, P. (1999). The boundaries of financial reporting and how to extend them. *The Journal of Accounting Research*, **37** (2): 353-385.
- [52] Lev, B. (2001). *Intangibles: Management, measurement and reporting*. Washington DC: Brookings Institute Press.
- [53] Libai, B., Narayandas, D. and Humby, C. (2002). Toward individual customer profitability model: A segment-based approach. *Journal of Services Research*, **5** (1): 69-76.
- [54] Linoff, S. G. (2004). Survival data mining for customer insight. [available at: <http://www.data-miners.com/links.htm>] (read in December 2004).
- [55] Lu, J. (2003). Modeling customer lifetime value using survival analysis. *SUGI*, (28).

- [56] Mani, D. R., Drew, J., Betz, A. and Datta, P. (1999). Statistics and data mining techniques for lifetime value modeling. *Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining*. Page 94-103. San Diego, California, United States.
- [57] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**:163-170
- [58] Microsoft Corporation. (2003). Microsoft Office Excel 2003. Redmond: Microsoft Corporation.
- [59] Moeschberger, M. L. and Klein, J. P. (1985). A Comparison of Several Methods of Estimating the Survival Function When There is Extreme Right Censoring. *Biometrics*, **41** (1): 253-259
- [60] Mohammed, Z. and Kotze, D. (2005). Survival data mining in the telecommunications industries: usefulness and complications, Data mining, text mining and their business applications, *Wessex Institute of Technology Transaction on Information and Communication Technologies*, Volume **35**: 505-512.
- [61] Mohammed, Z., Maritz, J. S. and Kotze, D. (2007a). Customer survival time in subscription-based businesses (case of Internet service providers). Data mining, text mining and their business applications, *Wessex Institute of Technology Transaction on Information and Communication Technologies*, Volume **38**: 303-310.
- [62] Mohammed, Z., Maritz, J. S. and Kotze, D. (2007b). Estimation of the customers' mean survival time in subscription-based businesses. Data mining, text mining and their business applications, *Wessex Institute of Technology Transaction on Information and Communication Technologies*, Volume **38**: 285-292.
- [63] Nelson, W. (1972). Theory and application of hazard plotting for censored failure data. *Technometrics*, **14** (4): 945-966.

- [64] Norris, F. (2001). Seeking ways to value intangible assets. *The New York Times*, 22 May.
- [65] Oakes, D. (2001). Biometrika centenary: Survival analysis. *Biometrika*, **88** (1): 99-142.
- [66] Octave Version 3.0 (2007). [<http://www.octave.org/>].
- [67] Oehlert, G. W. (1992). A note on the delta method. *Journal of American Statistical Association*, **46** (1): 27-29.
- [68] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariance test procedures (with discussion). *Journal of Royal Statistical Association*, **135** (2): 185-206.
- [69] Pfeifer, P. E., Haskins, M. E. and Conroy, R. M. (2005). Customer lifetime value, customer profitability and treatment of acquisition spending. *Journal of Managerial Issues*, **17** (1): 11-25.
- [70] Raab, D. (2007). Marketing systems: justifying marketing system investment. *DM Review Magazine*, June.
- [71] Reinartz, W. J. and Kumar, V. (2000). On the profitability of long lifetime customers: an empirical investigation and implications for marketing. *Journal of Marketing*, **64** (4): 17-35.
- [72] Roberts, J. H. (2000). Developing new rules for new markets. *Journal of the Academy of Marketing Science*, **28** (1): 31-44.
- [73] Rosset, S., Neumann, E., Eick, U., Vatnik, N. and Idan, Y. (2002). Customer lifetime value modeling and its use for customer retention planning. *Proceedings of eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*. Page 332-340. Edmonton, Alberta, Canada.
- [74] Rugimbana, R. (2007). Youth based segmentation in the Malaysian retail banking sector. *International Journal of Bank Marketing*, **25** (1): 6-21.

- [75] Rust, R., Zeithaml, V. and Lemon, K. N. (2000). *Driving customer equity: how customer lifetime value is reshaping corporate strategy*. Free Press: New York.
- [76] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69** (1): 239-241.
- [77] Shah, D., Rust, R. T., Parasuraman, A., Staelin, R. and Day, G. S. (2006). The path to customer centricity. *Journal of Services Research*, **9** (2): 113-124.
- [78] SPSS Inc. (2003). SPSS for Windows 13.0. Chicago: SPSS Inc.
- [79] SPSS Inc. (2005). SPSS for Windows 14.0. Chicago: SPSS Inc.
- [80] SPSS Inc. (2006). SPSS for Windows 15.0. Chicago: SPSS Inc.
- [81] Stata Corporation (2003). Stata release 8.0 manual: *Survival analysis and epidemiological tables*. Texas: Stata Press Publications.
- [82] Stata Corporation (2003). *Stata Statistical Software*: Release 8.0. Texas: Stata Corporation.
- [83] Therneau, T. M., Grambsch, P. M. (2001). *Modeling survival data: Extending the Cox model*, second edition. New York: Springer (this is a book).
- [84] Thomas, J. S. (2001). A methodology to link customer acquisition to customer retention. *Journal of Marketing Research*, **38** (2): 262-268.
- [85] Venkatesan, R. and Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, **68** (4): 106-125.
- [86] Walters, S. J. (2001). What is a Cox model? *Hayward Medical Communications, What is ? Series*, **1** (10).
- [87] Wangenheim, F. V. (2005). Postswitching negative word of mouth. *Journal of Service Research*, **8** (1): 67-78.


- [88] Weinstein, A. (2002). Customer-specific strategies: Customer retention: a usage segmentation and customer value approach. *Journal of Targeting, Measurement and Analysis for Marketing*, **10** (3): 259-268.
- [89] Weinstein, A. (2004). *Handbook of Market Segmentation*, third edition. Haworth Press: New York.
- [90] Wind, Y. (1978). Issues and advances in segmentation research. *Journal of Marketing Research*, **15** (3): 317-337.
- [91] wxMaxima Version 0.7.0 (2007). [<http://sourceforge.net/projects/wxmaxima/>].
- [92] Zeithaml, V. A., Bolton, R. N., Deighton, J., Keiningham, T. L., Lemon, K. N. and Petersen, J. A. (2006). Forward-looking focus: can firms have an adaptive foresight? *Journal of Service Research*, **9** (2): 168-183.



Appendix A

Octave code

A.1 The main file



```
clear
Data = load("DataFile.txt") ;
%Data is an m by four matrix of the data set;
global m y s a b t COV;
m = length(Data) ;
t = Data(:, 1); n = Data(:, 2) ; d = Data(:, 3) ; y = Data(:, 4) ;
%t is an m by one vector of time points where cancellation of service occurred;
%n is an m by one vector of number of customers at risk at each time point given in
the vector t;
%d is an m by one vector of number of customers who cancelled the service at each
time point given in the vector t;
%y is an m by one vector of probability of surviving beyond each time point given in
the vector t;
CovY = zeros(m, m) ;
for i = 1 : m
for j = 1 : m
for k = 1 : min(i, j)
```

```

CovY(i, j) = CovY(i, j) + d(k)/((n(k)-d(k))*n(k)) ;
end
CovY(i, j) = y(i)*y(j) * CovY(i, j) ;
end
end
s = 50 ;
SubCovY = CovY(s:end, s:end) ;
s is  $\tau_0$ ;
%CovY is the variance-covariance matrix of the elements in the vector y;
%SubCovY is the variance-covariance matrix of the elements of in the vector that been
used to estimate the parameters alpha and beta of the extrapolation function;
a = 0.829 ;
b = 0.011 ;
%a is the estimate of the parameter alpha in the extrapolation function;
%b is the estimate of the parameter beta in the extrapolation function;
G(s) = g(a, b, t(end), t(s), t(s)+0.1);
GA(s) = ga(a, b, t(end), t(s), t(s)+0.1);
GAA(s) = gaa(a, b, t(end), t(s), t(s)+0.1);
GAB(s) = gab(a, b, t(end), t(s), t(s)+0.1) ;
GB(s) = gb(a, b, t(end), t(s), t(s)+0.1);
GBB(s) = gbb(a, b, t(end), t(s), t(s)+0.1);
for i = s+1 : m-1
G(i) = g(a, b, t(end), t(s), t(i));
GA(i) = ga(a, b, t(end), t(s), t(i));
GAA(i) = gaa(a, b, t(end), t(s), t(i));
GAB(i) = gab(a, b, t(end), t(s), t(i));
GB(i) = gb(a, b, t(end), t(s), t(i));
GBB(i) = gbb(a, b, t(end), t(s), t(i));
end

```

```

%G is the function defined in equation;
%GA is the derivative of G with respect to alpha;
%GAA is the second derivative of G with respect to alpha;
%GB is the derivative of G with respect to betea;
%GBB is the second derivative of G with respect to beta;
%GAB is the partial derivatives of G with respect to alpha and beta;
bm1 = 0;
bm2 = 0;
for i = s : m-1
    B(1, i-s+1) = y(m)*GA(i);
    B(2, i-s+1) = y(m)*GB(i);
    B(3, i-s+1) = 0;
    bm1 = bm1 + B(1, i-s+1)-2*y(m)*G(i)*GA(i);
    bm2 = bm2 + B(2, i-s+1)-2*y(m)*G(i)*GB(i);
end
B(1, m-s+1) = bm1;
B(2, m-s+1) = bm2;
B(3, m-s+1) = 1;
Oaa = 0;
Oab = 0;
Obb = 0;
for i = s : m-1
    Oaa = Oaa + y(m) * ((y(i) - y(m) * G(i)) * GAA(i) - y(m) * GA(i)^2);
    Oab = Oab + y(m) * ((y(i) - y(m) * G(i)) * GAB(i) - y(m) * GA(i) * GB(i));
    Obb = Obb + y(m) * ((y(i) - y(m) * G(i)) * GBB(i) - y(m) * GB(i)^2);
end
Oaa = -2*Oaa;
Oab = -2*Oab;
Obb = -2*Obb;

```

```

%Oaa is the second derivative of the function omega defined in equation with respect
to alpha;
%Oab is the partial derivatives of the function omega defined in equation with respect
to alpha and beta;
%Obb is the second derivative of the function omega defined in equation with respect
to beta;
D = [Oaa Oab; Oab Obb];
INVD = inv(D);
DO = [INVD(1, 1) INVD(1, 2) 0; INVD(2, 1) INVD(2, 2) 0; 0 0 1];
COV = DO*B*SubCovY*B'*DO';
Delta = 12;
tau = 102;
d = 0.10;
alpha = zeros(Delta, 1);
Z = zeros(Delta, 3);
for i = 1 : Delta
    alpha(i) = 1/(1 + d/12)i;
    t1 = tau+i;
    Z(i, :) = Jac(t1);
end
S = Z*COV*Z';
MM = 100;
 $VarCLV = MM^2 * \alpha' * S * \alpha$ ;
survival=zeros(Delta, 1);
for i=1:Delta
    survival(i) = y(m) * g(a, b, t(end), t(s), tau + i);
end
CLV = MM * alpha' * survival;
%Delta is the time horizon in months at which projection has to be made;

```



%MM is the expected monthly from a customer;
 %survival(i) is the survival probabilities calculated from the extrapolation function at time i
 %VarCLV is the variance of customer lifetime value;
 %CLV is the customer lifetime value

A.2 Auxiliary functions

g.m

%This function evaluates the expression g as defined in 5.4.1. function y = g(a, b, tau, tau0, t)

$y = \exp(b * (\tau - \tau_0)^a - b * (t - \tau_0)^a);$

ga.m

%This function evaluates the first derivative of g with respect to alpha.

function y = ga(a, b, tau, tau0, t)

$y = \exp(b * (\tau - \tau_0)^a - b * (t - \tau_0)^a) * (b * (\tau - \tau_0)^a * \log(\tau - \tau_0) - b * (t - \tau_0)^a * \log(t - \tau_0));$

gaa.m

%This function evaluates the second derivative of g with respect to alpha.

function y = gaa(a, b, tau, tau0, t)

$os = b * (\tau - \tau_0)^a * \log(\tau - \tau_0) - b * (t - \tau_0)^a * \log(t - \tau_0);$

$y = g(a, b, \tau, \tau_0, t) * ((b * (\tau - \tau_0)^a * (\log(\tau - \tau_0))^2 - b * (t - \tau_0)^a * (\log(t - \tau_0))^2) + os^2);$

gb.m

%This function evaluates the first derivative of g with respect to beta.

```
function y = gb(a, b, tau, tau0, t)
y = exp(b * (tau - tau0)a - b * (t - tau0)a) * ((tau - tau0)a - (t - tau0)a);
```

gbb.m

%This function evaluates the second derivative of g with respect to beta.

```
function y = gbb(a, b, tau, tau0, t)
y = g(a, b, tau, tau0, t) * ((tau - tau0)a - (t - tau0)b);
```

gab.m

%This function evaluates the partial second derivative of g with respect to alpha and beta.

```
function y = gab(a, b, tau, tau0, t)
y = g(a, b, tau, tau0, t) * ((tau - tau0)a * log(tau - tau0) - (t - tau0)a * log(t - tau0)) * (1 + b);
```

Oaa.m

%This function evaluates the second derivative of omega with respect to alpha (omega is defined in 5.4.2).

```
function y = Oaa(a, b, tau, tau0, t)
```

V.m

%This function returns the variance of the extrapolated survival probability.

```
function v = V(t1)
global m y s a b;
y1 = y(m) * g(a, b, t(end), t(s), t1) * ga(a, b, t(end), t(s), t1);
y2 = y(m) * g(a, b, t(end), t(s), t1) * gb(a, b, t(end), t(s), t1);
y3 = g(a, b, t(end), t(s), t1);
Y = [y1; y2; y3];
v = Y' * COV * Y;
```

Var.m

%This function returns the standard error of the extrapolated survival probability.

```
function v = SE(t1)
global m y s a b t COV;
y1 = y(m) * g(a, b, t(end), t(s), t1) * ga(a, b, t(end), t(s), t1);
y2 = y(m) * g(a, b, t(end), t(s), t1) * gb(a, b, t(end), t(s), t1);
y3 = g(a, b, t(end), t(s), t1);
Y = [y1; y2; y3];
v = sqrt(Y' * COV * Y);
```

S.m

%This function calculates the confidence limits of the extrapolated survival probability.

```
function [L, P, U, se] = S(t2)
global y m a b t s
P = y(m) * g(a, b, t(end), t(s), t2);
se = SE(t2);
L = P - 1.96 * se;
U = P + 1.96 * se;
Endfunction
```

% The following function evaluates the derivative of the extrapolation function with respect to alpha, beta and y_m .

```
function M = Jac(t1)
global y m s a b t COV;
y1 = y(m) * ga(a, b, t(end), t(s), t1);
y2 = y(m) * gb(a, b, t(end), t(s), t1);
y3 = g(a, b, t(end), t(s), t1);
M = [y1 y2 y3];
```